CCSBT-ESC/2508/38
(ESC Agenda item 7.1)

Environment
www.csiro.au

Robustness of the CCSBT CKMR data

Rich Hillary

CSIRO Environment
Battery Point, Hobart 7000, Tasmania, Australia.

Copyright and disclaimer

Important disclaimer

Contents

Abstract

Given concerns raised in paper [1] about the robustness of the CCSBT CKMR data we provide a more detailed explanation of exactly what steps are taken to ensure that these data are indeed robust. We also provide additional simulated data examples to show why some of the issues raised in both [1] and the motivating paper [2] are either mistaken, or not relevant to the SBT case. The prime motivation of the kin-finding pipeline for SBT is to produce reliably identified specific kin pairs combined with a robust understanding of the false negative properties arising from the strict false positive criteria used to identify the kin. We conclude that the current kin finding pipeline is demonstrably robust and reliable, and that the intended next-generation augmentations to that pipeline will serve to ensure that robustness as the overall data set continues to grow.

1 Background

The concept of Close-Kin Mark-Recapture (CKMR) was first truly implemented for SBT, with both the Parent-Offspring (POP) and Half-Sibling (HSP) pairs fully integrated into the stock assessment and the Cape Town Procedure. Their centrality to the CCSBT science process serves to emphasise how important it is to have a robust and reliable process of data collection, genotyping, quality control and kin finding.

The primary motivating factors behind the CCSBT kin-finding process are:

- From the point of sampling extreme levels of care are taken to ensure high quality DNA is collected and sent for genotyping. Only samples that score highly are used and the genotyping errors of the samples are calculated and used in the QC process. The genotyped samples are of high quality with good read-depth and very low incidence of missing data
- An intensive Quality Control (QC) process that removes atypical loci, samples with properties that deviate from known statistical ranges/properties, and samples that are thought to be contaminated (as these look like related animals)
- Once all the approved loci and samples pass the QC process and the kin statistics for all the comparisons are computed, we calculate the key statistical features of the kin-finding statistics (e.g. where unrelated/less related comparisons are distributed, relevant false positive cut-offs and resultant false negative probabilities)
- Once we have an unambiguous set of specific kin pairs we then check the sample metadata to further check the bioinformatics results and be as certain as we can that as to the identified kin type (e.g. a POP where the adult was less than 5 years old when the juvenile was born is very suspect and could be an FSP)

One thing we are **not** trying to do is assign a probability of kinship to each pair being compared. This is something that `COLONY` attempts to do but it is worth including more detail as to why we think this is not a good strategy in relation to CKMR as the CCSBT implements it. To have a (posterior) probability of kinship for a given comparison we would need a prior probability of kinship, to be updated by a likelihood of kinship using the genotyping data. This prior probability, however, depends on both the kin type *and* the covariates of both the samples. For example, for an HSP even with a stable adult population size the prior probability of being an HSP will decrease the further apart in time the birth years of the samples being compared are due to mortality. In fact, calculating this prior probability of kinship is exactly what we are doing in the likelihood functions for both POPs and HSPs in the CCSBT OM. These probabilities get rapidly

more complicated to do as the kin type under consideration gets more distant (e.g. third and fourth order kin). This is why the primary focus of the CCSBT CKMR implementation is to find an unambiguous set of first (POPs, FSPs) and second (HSPs) order kin, not have to deal with pedigrees and construct demographic priors of ever increasing complexity for more distant kin that cannot be totally removed from the samples.

2 Answers to specific questions raised by Japan

In this section we try and answer the specific questions raised in paper [1]:

2.1 Are PLOD approaches biased?

The specifics of exactly what is suggested to be biased is unclear in [2]. The first mention in relation to the genotyping and kin-finding is on page 7 and relates to the Pacific bluefin results in terms of missing genotype results. This does not apply to the SBT case given (i) the intensive QC process; and (ii) the CCSBT use a very different genotyping platform which produces consistent high read-depth, and both a very low incidence of missing data and robustly identified true null alleles. The next mention of bias appears on page 9 and concerns potential bias of the PLOD approach relative to likelihood inference in `COLONY` or in the `fraRF` approach outlined, though not clearly defined, in [2]. Specifically it appears to be an assertion that different numbers of markers means a PLOD approach must be biased. This is, if we understand the assertion as it is stated in the paper, incorrect. Firstly, for SBT all the fish are genotyped at the same SNPs - there is no difference in the SNPs being used in all the comparisons so there could not be any bias appearing due to this issue (and the bias would only appear in a specific case - see below). The authors seem unclear on what the PLOD approach really is doing and how it works. The PLOD means Pseudo-Log Odds and is an approximation to the true likelihood ratio of comparative kin types. The “pseudo” part of the definition means that, because the approach ignores genetic linkage of SNPs on particular chromosomes, although the expected estimate of the PLOD for a given comparison will be correct, the variance will be under-estimated. Hence, we do not use the PLOD to try and calculate the variance in the expected PLOD of specific kin types. So, in the way it is being used (i.e. only the estimated PLOD not its variance) it will not be biased in the way being intimated in [2]. This is something we will demonstrate later on with our simulations where genetic linkage is explicitly included but the PLODs still appear centered around exactly where the PLOD approach suggests that they would. The only bias using differential numbers of markers could cause in the PLOD approach is where one doesn’t recalculate both the expected PLODs for the kin types as well as their false positive and negative criteria. We would therefore conclude that it is highly unlikely that we have any bias in the identified kin pairs caused specifically by using the PLOD approach.

The PLOD issue only relates (directly) to HSPs even though POPs, FSPs, and HSPs are discussed in [2]. Prior to the introduction of SNPs in 2015 microsatellite loci were used and specifically for POP identification using the principle of Mendelian exclusion: an offspring must share an allele with one of its parents. Those that share at all locations must be POPs; those that do not are not. For the SNP data the presence of null alleles complicates the exclusion issue because called genotypes of true POPs might mismatch because of genotyping error (or, less likely, mutation). The CCSBT kin finding pipeline uses a weighted exclusion statistic to deal with this issue that maintains the power of Mendelian exclusion to find POPs, but accounts for the locus-specific probability of genotyping errors causing a mismatch. We could use the PLOD approach to identify the POPs but have maintained the weighted exclusion approach for SBT.

Where the PLOD approach does become useful is for cases such as FSPs that may look like being close to being POPs - they are both first order kin pairs and FSPs are most likely to score higher exclusion statistics than HSPs and lower order kin types. This is something that is also done in the CCSBT kin finding pipeline. From Figure 2 in [3] any comparisons that are close to the false-positive cut-off and have covariates that indicate they could be FSPs can then be tested with a specific POP *vs.* FSP PLOD test that very clear shows which one the comparison really is.

2.2 Degraded samples and the QC process

As mentioned in the background section there is a rigorous process of quality control that is undertaken across multiple different dimensions of concern. First and foremost, degraded samples that have too little DNA or give poor reads are not used. Later on there is also a specific criterion used to detect an over-representation of null alleles that may indicate degraded samples and those samples are removed from subsequent kinship analyses. As the data collection and processing pipeline has improved over time the proportion of retained samples has, as noted in [1], improved greatly. However, the main point to make is the following: what matters most is being sure of the sample's quality and that we retain *enough* samples to be sure of obtaining an informative amount of kin pairs. The fact that a certain amount of samples are removed is a sign of the strictness of the QC process, not an indication of the quality of the retained samples. Given there was no request from members prior to paper [1] being submitted it was not possible to elaborate on the specific details of the QC process in the usual CKMR paper [3]. If members require additional technical detail we are happy to accommodate this but note that traditionally this group has noted its limitations in relation to being able to adequately discuss these details. The are several statistically detailed steps to the QC process that would need explaining given the request from Japan.

3 What really matters for kin finding for CKMR

In this section we outline in more detail exactly what steps are taken to robustly identify HSPs, and why the specific PLOD approach as used in the CCSBT kin finding pipeline [3] is highly likely to produce a robust data set that we can be confident in. In [2] the authors appear to suggest that a particular method is more reliable if it produces less, or even zero, false positives using simulation methods. This is not in fact what matters for the CCSBT CKMR modelling - what matters most is that:

1. We are able to estimate a strict enough false-positive threshold in our kin finding statistic so that our data set has a high probability of not including more distant kin pairs
2. We have a robust statistical understanding of the resultant false-negative properties of the remaining kin pairs that are above the false-positive threshold

In reality, given enough comparisons or lower numbers of SNPs there are **no** kin finding algorithms that will be able to exclude **all** false-positives. In our view, any major CKMR project on a large population will face this issue - the PBF case outlined in [2] may be an outlier in that regard given the current size of the population, and the number of samples and loci used in [2] but this does not apply to SBT as we already know [3]. The authors of [2] in fact specifically raise the very issue that makes the PLOD approach preferable to either the `COLONY` or `fraRF` approach on page 9 of the paper:

> *"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"*.

Here it is important to note that the authors of [2] did not use the software packge used in the CCSBT kin finding pipeline. The `kinference` package used for CCSBT kin finding, while using several similar features to `CKMRsim` such as PLOD statistics for various kin types, is designed to robustly estimate the key false positive and negative properties that are so important to the process. The current HSP false-positive/negative process [3] works as follows:

1. Mixture distributions are used to estimate the PLOD distribution of third (HTPs) and fourth (HCPs) order kin
2. A suitable upper threshold to the PLOD is defined whereby we would expect less than 1 true lesser kin (HTP, HCP) to appear within the true HSPs
3. Using the true HSPs that appear above the expected HSP PLOD we calculate a one-sided variance estimate of the PLOD distribution for HSPs
4. Given this variance estimate we then calculate the fraction of true HSPs we expect to see above the false-positive threshold
5. This is the false-positive retention probability then used in the HSP CKMR likelihood in the OM (π^{η})

4 Why the PLOD approach works even with genetic linkage

More recently, with the availability of whole-of-genome sequencing at low cost it as been possible to get the genome of many different species. This gives us the ability to include the genetic linkage of the various loci in the kin finding process thereby increasing its statistical power [4]. CSIRO has mentioned over the last few years its intention to begin adapting the kin finding algorithms to accommodate genetic linkage now a genome is available. What is required is a linkage map that tells us specifically how the loci on a given chromosome co-inherit when parents produce offspring. This actually exists for Pacific bluefin tuna [5] but we have not yet obtained one for SBT - there are several ways to do this but PBF was able to find offspring with known mother *and* father due to the closed breeding cycle which makes the linkage map estimable in a relatively straight forward manner. The PBF linkage map maybe a useful starting point for SBT and future work is under way to develop this linkage map for future use.

In this section we demonstrate why the PLOD approach used in the CCSBT is workable even when we have different linkage dynamics via a simulation approach using the CCSBT genotyping data [3] (allele frequencies, genotyping error) and three linkage scenarios (low, medium and high) whereby the approximately 1,500 SNPs are evenly distributed (low), slightly clustered (medium) and highly dense and co-inherited (high) across 22 chromosomes. The linkage is simulated via the recombination factor, θ_l, at each locus [4] with $\theta_l = 0$ meaning adjacent loci l and $l + 1$ are always co-inherited, and $\theta_l = 0.5$ meaning unlinked loci whereby a recombination event is a 50/50 outcome due to random assortment. It is assumed that when a locus l is at the end of one chromosome and $l + 1$ is at the top of another then $\theta_l = 0.5$. The genotypes and IBD state (0, 1 or 2 in terms of alleles shared by descent) of each locus form a Hidden Markov Model (HMM) for the key kin types we consider in CKMR (POPs, FSPs, HSPs, HTPs, HCPs) [4, 6]. Specifically, the IBD state is the hidden state, and the genotype is the observable process. The locus-specific transition matrix that defines the HMM is defined by the recombination

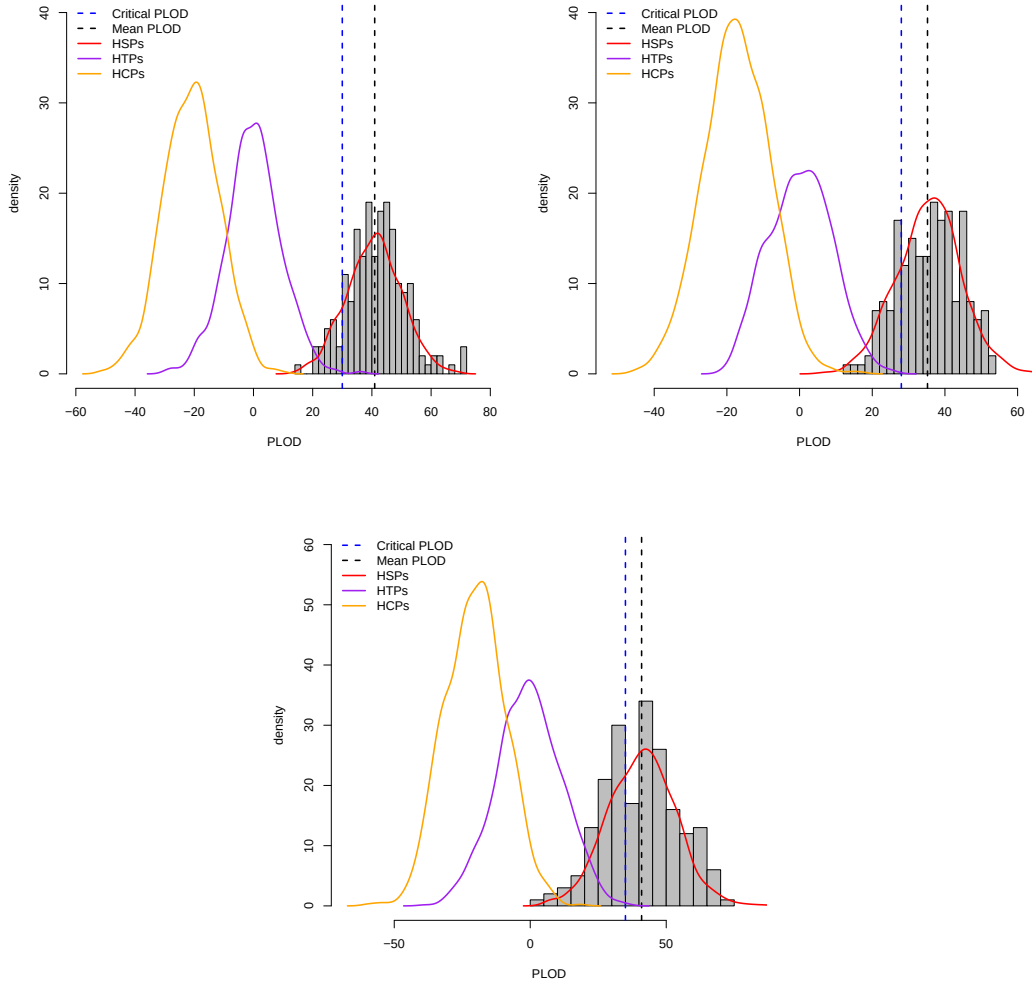

Figure 4.1: *HSP PLOD summary for the low (top left), medium (top right), and high (bottom) linkage scenarios. HSP PLOD distributions for true HCPs is orange, HTPs is purple, HSPs is red. The critical false-positive is the dashed blue line, and the expected HSP PLOD is the black dashed line. The histograms in each plot are the 200 simulated HSPs.*

factor θ_l and the specific kin type of interest [7]. It is anticipated that when we have the genetic linkage map for SBT we will use these HMM algorithms to replace the current PLOD approach with a fully likelihood based kin finding algorithm but for now we use them in simulation mode for instructive purposes in this paper.

For the simulations we simulate 200 (close to the current SBT number of around 220) true HSPs for each of the linkage scenarios. The HSP PLOD distribution for the case where animals are true HTPs (third-order) and HCPs (fourth-order) is then calculated so the critical false-positive PLOD threshold can be estimated. In all cases this is driven by the closer true HTP PLOD distribution (purple line in Figure 4.1). In Figure 1 the red distribution is the true HSP PLOD distribution (which we don't know) and the histogram is the distribution of 200 HSPs. In each case we then use the HSPs above the expected HSP PLOD (dotted black line) to estimate the variance of the HSP PLOD distribution. Assuming a normal distribution we then calculate the probability that a true HSP would appear below the critical false-positive PLOD threshold (this is

the false-negative probability). Our set of HSPs that we use for CKMR analyses are the observed HSPs with a PLOD above the threshold, and 1 minus the false-negative probability is our HSP retention probability used in the CKMR likelihood.

We did not have time to do a full simulation-evaluation of this process but even with three simple examples the following should be clear:

- We are highly unlikely to find true false-positive HSPs in our SBT samples given the strictness of the false-positive threshold calculation - irrespective of the strength of linkage
- In all cases we see that the true PLOD distribution and empirical distribution of HSPs (and, by implication the one-sided estimate of PLOD variance) show close similarity. For the three low/medium/high linkage scenarios the true false-negative probabilities (in terms of actual HSPs below the critical value) were 0.08, 0.2, and 0.33, respectively. The one-sided variance and normal approximation estimates were 0.05, 0.23 and 0.28

In the future, where we utilise the genetic linkage map and the more powerful HMM kin finding suite of algorithms we fully expect to get both better discrimination between the HSPs and their third and fourth-order nearby kin, and a simulation based understanding of their various LOD (not PLOD) distributions. This means we can move away from the mixture distribution based false-positive threshold calculations and the one-sided variance false-negative normal approximations. We can calculate these values directly via efficient Monte Carlo simulation and be even more accurate with our false-positive and related false-negative variables.

5 Discussion

Japan raised a number of questions in [1], motivated by some of the points made in [2] relating to kin finding in Pacific Bluefin tuna. In this paper we have systematically addressed the specific questions raised, highlighted issues in [2] that are either not applicable to SBT or confuse aspects of the general approach used in CCSBT, and using simulated data to highlight exactly how the CCSBT process works and is likely to be robust.

In future, especially given we expect some technical and detailed changes to the future kin finding program, it may be useful for more detailed presentations of the whole kin finding platform. However, the view of this paper is that the current CCSBT kin finding process has the most developed and strict QC process, and that the existing CKMR data sets (both POPs and HSPs) are very robust and usable in they way the CCSBT is currently using them.

6 Acknowledgements

This work was funded by CSIRO and the Australian Fisheries Management Authority.

References

[1] Itoh, T. (2025) Questions regarding CCSBT's close-kin mark-recapture work from Japan. *CCSBT–ESC/2508/26*

[2] Tsukahara, Y., Nakamichi, R., Matsuura, A., Akita, T., Fujiwara, A., and Suzuki, N. (2025) Comparison of kinship-identification methods for robust stock assessment using close-kin mark–recapture data for Pacific bluefin tuna. *Pop. Ecol.* 1–13.

[3] Farley, J., Eveson, P., and Gunasekera, R. (2025) Update on the SBT close-kin tissue sampling, processing and kin-finding 2025. *CCSBT–ESC/2508/8*.

[4] Boehnke, M., and Cox, N.J. (1997) Accurate inference of relationships in Sib-Pair linkage studies. *Am. J. Hum. Genet.* **61**: 423–429

[5] Uchino, T. *et al.* (2018) Genotyping-by-sequencing for construction of a new genetic linkage map and QTL analysis of growth-related traits in Pacific bluefin tuna. *Aquac. Res.* **49**: 1293–1301.

[6] McPeek, M.S., and Sun. L. (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.* **66**: 1076–1094

[7] Dudiot, S. (2003) IBD configuration transition matrices and linkage score tests for unilineal relative pairs. *University of California Berkeley technical paper*

CONTACT US

t 1300 363 400
+61 3 9545 2176
e csiroenquiries@csiro.au
w www.csiro.au

WE DO THE EXTRAORDINARY EVERY DAY

We innovate for tomorrow and help improve today for our customers, all Australians and the world.

Our innovations contribute billions of dollars to the Australian economy every year. As the largest patent holder in the nation, our vast wealth of intellectual property has led to more than 150 spin-off companies.

With more than 5,000 experts and a burning desire to get things done, we are Australias catalyst for innovation.

WE IMAGINE. WE COLLABORATE. WE INNOVATE.