GENOTYPING ISSUES FOR CKMR ON ATLANTIC BLUEFIN TUNA

MARK BRAVINGTON, ROBIN THOMSON, CAMPBELL DAVIES: CSIRO, 15 FEB 2017

1. Introduction

Davies et al. (2017) present the broad outline of a possible Close-Kin Mark-Recapture (CKMR) program for EABFT, showing what types of sampling and modelling could and could not work. The *potential* benefits to assessment and management are considerable and also unobtainable through any other type of data, in particular by providing objective information on:

- true fecundity-at-size;
- subtle stock structure issues within the Eastern Atlantic;
- absolute abundance, now, for the spawning stock(s), without relying on contentious past or present CPUE data;
- monitoring method that can robustly track future changes in spawning-stock.

To realize those benefits, at least two further preliminary steps would be needed before starting a full program:

- a more detailed consideration of practical sampling options, including options for staged designs, requirements for ancillary data, and specificity about locations;
- agreement on a genotyping method for kin-finding.

This document addresses the second step. CKMR requires accurate and thorough genotyping in order to identify pairs of kin. Picking a good method is crucial, but not easy: there are many different modern methods of genotyping; the investment in time and money required to master any one method is substantial; and there are issues specific to CKMR which are far from obvious. This report explores some of the latter, drawing on our experience at CSIRO of CKMR projects for several species. We put a lot of effort into choosing a modern genotyping method for those projects, and while we are very satisfied with the results, there are alternative methods which could be worth considering. One conclusion of this report is that the whole area is so technical that a well-founded decision is unlikely without a dedicated workshop.

To allow effective comparison of candidate genotyping methods during the workshop, these methods should be applied to sample DNA beforehand. Samples for ABFT CKMR (both the Eastern and Western populations) might be collected by several countries or organizations. Whoever actually does the genotyping, it is essential that all samples are genotyped at the same loci, so that kin-comparisons can be made across regions and time. In practice, this also means that all genotyping should be done the same way, to avoid unnecessary complications at the stage of pairwise kinship comparison [1]. The number of loci genotyped, their information content re kinship (see later), and the quality of genotyping, must be good enough to find Half-Sibling Pairs (HSPs). This is a demanding task which is entirely possible with the best modern methods, but entirely impossible if an inadequate method is used. As discussed in Davies et al. (2017), HSPs as well as Parent-Offspring Pairs (POPs) are essential for CKMR for EBFT, for several reasons:

(1) HSPs from *different* cohorts give information on spawning site fidelity in adults, which is relevant to CKMR (and management) for EABFT and unobtainable any other way;
(2) without HSPs from *different* cohorts, it is impossible to separately estimate (adult) mortality and selectivity in a robust fashion, even if POPs are available;[2]
(3) when sampling young juveniles from the *same* cohort, there is a possibility that "litter-mates" from the same spawning event may be disproportionately likely to be caught together (full-sibs and/or half-sibs). It is essential to estimate this proportion, which hopefully will be very small; but if it is not small, then it must be allowed for in subsequent steps of CKMR, otherwise variance estimates will be too optimistic.

The need for HSPs does impose some quite stringent demands on the genotyping method. Since any method capable of reliably finding HSPs will have no difficulty at all identifying POPs (and full-sibling pairs), here we focus on the challenge posed by HSPs.

In the rest of this document, Section 2 is a technical but informal summary of the statistical aspects of HSP-finding for CKMR, as driven by basic genetics and genotyping. It is based on our experience with SBT and three shark species using variants of ddRAD-based Genotyping-By-Sequencing (GBS), but the statistical principles are quite general and should apply to almost any species and genotyping method. A briefer but more mathematical presentation may be found in section 5 of Bravington et al. (2016b), which also gives a statistical and theoretical overview of CKMR in general. In Section 3, we propose an outline for a genotyping workshop. Section 4 gives a summary and recommendations for the next steps, should ICCAT decide to proceed with CKMR for EABFT.

2. Principles of HSP-finding for CKMR

The target audience for this section consists of both geneticists and statisticians interested in issues around large-scale kin-identification for CKMR; as such, depending on who is reading, some parts may seem obvious and others incomprehensible. The first, theoretical, step is to identify some statistic (the one proposed here we call the "PLOD") can be calculated for each pair of fish using their genotypes at all loci, and which is likely to be large if the fish are Half-Siblings and small otherwise. The practical issue is then to ensure that enough loci are genotyped with enough accuracy so that, for most HSPs,

[1] There are numerous published examples; see Ellis et al. (2011) for a fairly recent case. The details depend completely on the genotyping method(s), but the whole unwelcome problem can be avoided by using one single method with a single service provider and fixed protocols.

[2] In theory, mortality and selectivity *can* be separately estimated in a stock assessment that has perfect data, no changes in catchability-at-age, full age information, etc. In practice these conditions rarely seem to be met, and ABFT is no exception. We were able to work around the issue for SBT only thanks to special circumstances.

their PLOD will be high enough to clearly separate them from all unrelated and less-closely-related pairs. With careful planning and preliminary information on the performance of a genotyping method, this can— and certainly should— be done *before* genotyping the bulk of the samples, so that there are no nasty surprises and no need to re-do any genotyping post hoc.

All the figures shown are schematic, in that the heights and widths and locations of the bumps have been chosen for clarity rather than taken from real data. However, the real examples that we have examined do follow these patterns. For the sake of clarity, we have glossed over many details and avoided discussing rare exceptions; for more depth, see e.g. Thompson (2000) and any number of introductory genetic textbooks and websites. In order to make this document reasonably self-contained, however, we have included some simplified explanations in footnotes and in Appendix A.

We focus here on the issues around HSP-finding *after* genotyping has been done. "Genotyping" here means the entire process that transforms raw genetic data into "the genotype"— the inferred pair of alleles for each sample at each locus[3]. For a SNP locus with just two alleles called A and B and no null-alleles (see 2.2), the possible genotypes are AB, AA, and BB. The nature of "raw genetic data" depends on the genotyping method; with GBS approaches, which we think are currently the only practical way to find HSPs for most species, the raw data consist of gigabytes of output on "reads" from sequencing. The details of genotyping are method-specific and largely irrelevant here, where the focus is on what to do with the genotypes after finding them, but it is important to be aware that there are always *many* steps involved in genotyping, which need to be carefully tailored to the specific genotyping method being used.

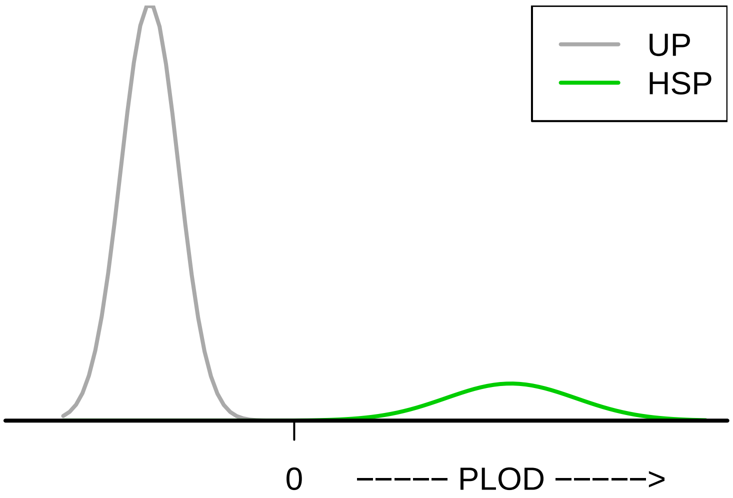
And even before genotyping and starting to look for HSPs, a great deal of method-specific background work is necessary in order to: select loci that are "valid", including but not limited to the elimination of paralogous loci; eliminate bad samples (e.g. contaminated DNA); confirm that genotype frequencies across samples are as they should be (according to Hardy Weinberg equilibrium); estimate genotyping error rates, etc. Again, these are method-specific details that are out-of-scope for this document. However, to evaluate any genotyping method proposed for CKMR for ABFT, summaries of these steps would need to be considered.

2.1. **Distinguishing half-siblings from unrelated pairs in large samples.** The primary question to consider is whether a pair of animals is a Half-Sibling Pair (HSP), or an Unrelated Pair (UP), under the assumption that exactly one of those two kinships is correct. Later in section 2.4, we discuss how other types of kin affect matters.

A statistically efficient way to summarize the genotypic evidence for HSPness/against UPness of a pair of fish, is via the PLOD (Pseudo-Log-ODds ratio, based on a log-likelihood-ratio familiar to geneticists), described in A.1 or Bravington et al. (2016b). This is a single positive-or-negative number which tends to be bigger for HSPs and smaller for UPs. Figure 1 shows an idealized histogram of PLOD values across a large number of pairwise comparisons in a large population, where each pair is either an UP or an HSP, and when a very large number of highly-informative loci are genotyped with good accuracy.

[3] "Genotyping" is also sometimes referred to as "calling" the genotype, and the word "genotype" sometimes also refers to the entire collection of genotypes across all loci for one individual sample.

Figure 1. Idealized separation of UPs from HSPs based on PLOD. No other kin present.



The figure illustrates several general points:

(1) The left-hand bump, from UPs, is bigger than the right-hand bump, from HSPs. In practice, the UP bump would actually be *several million* times larger than the HSP bump; HSPs arise when two animals share a parent, and if there are millions of possible parents, HSPs will be correspondingly rare. “Larger” means that both the height and the width would be proportionally greater.

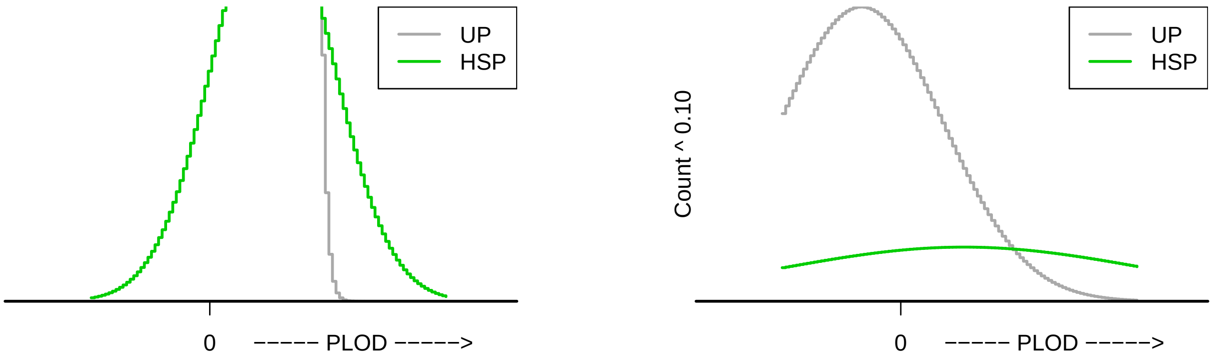
(2) The main reason that the HSP bump is wider than the UP bump, is variability in the *amount* of co-inherited DNA in different HSPs. There is a 50% probability that any single locus in an HSP will include one co-inherited copy; but DNA comes in strands, and physically-nearby loci will tend to share the same co-inheritance status— ie they are “linked” (sections A.2 and A.3). For close-kin such as HSPs, but not for UPs, this leads to lack of statistical independence between the per-locus components of the PLOD, and hence higher PLOD variance for HSPs, and hence a wider bump.

(3) In this idealized situation, it would be easy to distinguish HSPs from UPs, because there is clear separation between the bumps. Of course, in practice there is overlap between the two bumps and one does not know the “colour” of an given pair in the overlap region; the point is to classify each pair based on its PLOD.

The key to successful HSP-finding is to ensure that enough of the HSP bump on the right is clearly separated from the UP bump (and bumps from other less-close kin; section 2.4). One very important point is that the entire statistical distribution of the UP bump (height, location, and spread) can be predicted in advance, and also checked against the observed histogram, the vast majority of which will consist of UPs. For the HSP bump, the mean can be predicted, but not the variance (i.e. spread), because the extent of linkage cannot be known in advance; and not the height, because that is driven by the very population size that we are trying to estimate.

In practice, we (i) probably will not have, (ii) possibly cannot have, but (iii) fortunately do not actually need, enough genotyped loci to get such clear separation as Figure 1. To show what could go wrong with too few loci, Figure 2 zooms into the overlap region, where this time UPs are (as in reality) much more common than HSPs. The UP bump spreads well into the HSP bump; the nature of tail-probabilities, and the far greater number of UPs to HSPs, makes the UP bump seem more like a cliff (note that the histograms show expected values, so are much smoother than would really be seen). This situation would clearly not be satisfactory; there are only likely to be a few hundred true HSPs, and though a handful would still show up on the far right-hand side, the curves cannot clearly be separated, assumptions cannot be checked, and it would be impossible to estimate how many HSPs were lost if a PLOD threshold was set. Metaphorically speaking, the green hillock of truth has been buried under an avalanche from the grey mountain of irrelevance.

FIGURE 2. Poor separation with too few loci. Left-Hand Side is comparable with Figure 1 and shows what would be seen for an idealized large dataset; RHS shows the same data as LHS, but with almost-log-scale on the y-axis, in order to fit both UPs and HSPs.

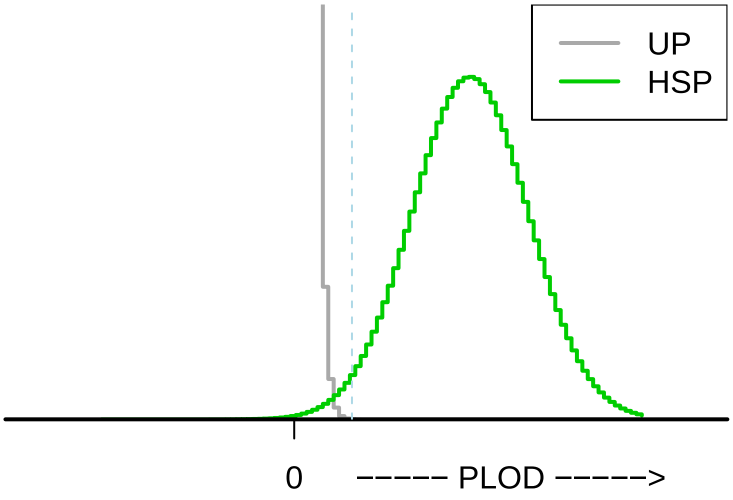


2.1.1. *A systematic algorithm.* By increasing the number of SNPs in Figure 2, we could eventually get reasonably clear separation (assuming the new SNPs are much like the old in terms of their power to detect HSPs), as in Figure 3. At that point, a statistically valid approach to finding HSPs is as follows:

(1) Check that the left (UP) bump closely coincides with the theoretical mean and variance for UPs. If not, then there is some problem with the loci/genotyping/modelling/assumptions, which needs to be fixed.

(2) If the UP bump has the correct mean and variance, then take some rough idea of the likely number of true HSPs and choose a threshold PLOD value η, high enough that the number of

FIGURE 3. Reasonable separation with more loci. Blue dashed line could be a reasonable threshold η.



UPs likely to have PLOD $> \eta$ is no more than a small proportion— say, well under 1%— of their total number. This ensures that false-positives from UPs will not be a problem. The likely number of true HSPs does not need to be precise, just order-of-magnitude; it would presumably come from a design study such as Davies et al. (2017), erring on the low side when in doubt. However, calculation of the tail probability for the UP bump needs some mathematical care, because the tail probability is so extreme, e.g. 10^{-6}; for example, a Normal approximation should not be reliable.

(3) Confirm that the mean PLOD for HSPs is about where it should be (sample size will be limited here, but *some* bump corresponding to HSPs should be clearly visible). If not, this again indicates that there is some underlying problem to fix, as in step 1.

(4) If η is above or even close to the mean PLOD for HSPs, then more loci and/or more accurate genotyping will be needed.

(5) Otherwise, estimate the PLOD variance for HSPs based only on those pairs with PLOD above the expected value for HSPs (see Bravington et al., 2016b). This automatically accounts for the effects of linkage.

(6) Using the variance from the previous step, estimate the probability that a true HSP will have PLOD $< \eta$, i.e. will be a false-negative. This probability is the key ingredient for developing an unbiased CKMR model based on "definite HSPs" (i.e. pairs with PLOD $> \eta$) with appropriate allowance for false-negatives.

(7) The threshold η may need to be refined to exclude more distant kin such as cousins; see section 2.4.

The actual number of loci needed for HSP-finding depends on the species, partly because the size of the UP bump relative to the HSP bump depends on the population size, and partly because the number of potential loci tends to vary between species, and it is easier to pick good loci if there are more to choose from. For SNPs, the theoretical minimum number of with maximally-informative loci needed may be somewhere around 1000–1200, though we have never tried to work it out exactly because in practice it is not easy to find maximally-informative loci (e.g. all with Minor Allele Frequency, MALF, ≈ 0.5). In practice, we have managed to get good separation of bumps for several species using around 1500 loci, but only after carefully selecting loci with high information content (e.g. MALF certainly >0.1 and preferably close to 0.5; low incidence of nulls, except when using the technique mentioned at the end of section 2.6; of course, basic QC such as Hardy-Weinberg checks must also be satisfied).

Given that so many SNPs are required for HSPs, targeted assays ("SNP chips") would appear to be out of the question financially; GBS seems the only affordable way. However, there are many different GBS methods, which vary in the the quality of genotyping, quantity of loci, and cost; different GBS methods are suitable for different purposes, and the suitability of a method for CKMR needs to be checked specifically.

2.2. **Null alleles.** Each sample has two copies of each locus, which may carry the same or different alleles. But sometimes genotyping will only reveal one copy (or none)— the "missing" copy is a null allele[4]. The presence of nulls can be inferred by first assuming they don't exist and then looking for a "heterozygote deficiency" amongst a large (say, 100+) sample of genotypes assuming Hardy-Weinberg equilibrium (HW), given the estimated allele frequencies under the HW assumption (although, notoriously, there are many other reasons why HW can fail, so those possibilities also need to be ruled out). Failing to deal with this would be a major QC error which would leave the UP and HSP bumps in badly the wrong place, and ruin the identification of HSPs. However, heterozygote deficiency is a routine QC check, and there is no technical difficulty in estimating the null frequency for each locus based on its "heterozygote deficiency". After adjusting the PLOD calculation accordingly, the UP and HSP bumps should end up in the right place. The main issue with nulls is the difficulty or impossibility (but see section 2.6) of distinguishing an "AO" genotype (one copy having allele A, the other copy having a null allele) from an "AA" (homozygote, i.e. two copies of A). Having to genotype such cases as "either AA or AO but not sure which" can substantially reduce the power of a locus for kin-finding, compared to another hypothetical locus with the same MALF but no nulls.

Whether a copy is read as a null depends on the genotyping method, and of course the species and loci involved. With microsatellites in SBT we found a low incidence (1%–3%) of nulls in most of the loci we used, rising to 15% in a few loci that we used later only in confirmation tests; however, we

[4] That definition is somewhat loose, but this is not the place to try to make it precise. There is a terminological minefield around nulls; for example, "random" genotyping error can also cause an allele that is "really" present, to be overlooked just by chance. Probably, the most useful distinction is between "heritable nulls" (which will repeatedly appear in repeat genotyping, and also manifest themselves in descendents) and nulls-through-error. In this document the focus is primarily on the former, but the two are not necessarily easy to tell apart.

deliberately tried to choose loci with nulls. With our current GBS methods, we have found nulls to be *very* common in tuna (and less common, though still prevalent, in sharks), and there simply aren't enough suitable null-free loci to avoid the null issue altogether. The good news is that, given good sequencing and careful genotyping methods, the presence of nulls can actually be turned from a curse into a blessing for kin-finding; see Section 2.6.

Null alleles are (in our experience) probably the least-ignorable unwelcome complication of genotyping, but they are certainly not the only one. Phenomena such as Copy-Number Variation and paralogous loci seem to be quite prevalent in GBS, but they really belong to the preliminary QC phase rather than the identification-of-HSPs which is the focus of this section.

2.3. **Genotyping errors.** With thousands of samples and thousands of loci, calling genotypes needs to be an automated process, and the data are not always unambiguous. In principle it might be possible to record specific loci/samples as "not sure" and thereby steer clear of errors. However, kin-finding is computationally demanding even with "definite" genotypes, and to avoid insurmountable computational complexity, we consider that it is much better to always call a single definite genotype at each fish and locus[5]. To deal with the inevitability that some calls will be wrong, overall probabilities of genotyping error[6] should be estimated, certainly using replicate samples and if possible also using known POPs[7].

Genotyping errors (unless negligibly small) have two possible effects on HSP-finding:

- If the errors are not allowed for when computing PLODs, the UP and HSP bumps will not fit the theory, and classification of HSPs will go wrong; this would be a serious problem.
- If errors are properly allowed for in the PLOD, then the UP- and HSP-bumps will overlap more than if no errors were present.

2.4. **Other kin.** There are three types of non-HSP kin to consider: more-related than HSP, equally-related (on average), and less-related. Here, "related" means "expected proportion of Identical-By-Descent (IBD) alleles"; random variability can affect the actual IBD proportion of a given pair, so occasionally a pair of cousins will share more genetic material than a pair of siblings.

More related:: Duplicate fish (and identical twins), POPs, and Full-Sibling-Pairs (FSPs) are immediately obvious in PLOD histograms as clusters or outliers way to the right. These kin-types can easily be distinguished by further checks, and errors are negligible given the number of loci required for HSPs. Note that FSPs of different ages should be very rare in typical long-lived teleost CKMR, since they require the same two adults to mate in different years. (Within-cohort

[5] For some loci, it may be preferable to group certain genotypes; e.g. if the alleles are A,B,O (null) as in section 2.2, then it *may* be best to assign each sample only to one of the 4 categories AB, AA/AO, BB/BO, OO rather than to try to separate AA from AO.

[6] Genotyping error rates will vary according to the locus and the specific genotype that is called, but (in our view) no attempt should be made to estimate sample-specific rates, because that makes kin-finding spectacularly more complicated. Certainly, some samples will have worse quality DNA than others, but the practical way to deal with that is to run QC at the sample level, rejecting entirely those samples that show (for example) an unlikely high or low proportion of heterozygotes (which can arise from contamination and from poor DNA extraction, respectively).

[7] In the early phases of CKMR for a new species, POPs will probably not be available.

FSPs are possible if larvae/juveniles from the same spawning event continue to associate until the point of capture; see Bravington et al., 2016b.)

Equally related:: HSPs have the same average relatedness as GGPs (Grandparent-Grandoffspring Pairs) and FTPs (Full-Thiatic-Pairs[8], i.e. Aunt/Nephew etc). These pairs cannot be directly distinguished from PLODs. In theory, there would be *some* prospect of distinguishment given "dense marker sets" (enormous numbers of SNPs) and a very good genome assembly, but— even if possible— that level of genotyping is both unaffordable and unnecessary for fish. FTPs should be very rare, basically for the same reason as FSPs; GGPs cannot arise if the animals-to-be-compared are lethally sampled before maturity (i.e. if HSP comparisons are restricted to immature fish, as proposed for EABFT).

Less related:: The next-most-related type of kin are HTPs (Half-Thiatic-Pairs, i.e. grandparent of one member is parent of the other) and after that HCPs (Half-Cousin-Pairs, i.e. one shared grandparent). HTPs are rarer than HSPs[9], while HCPs are commoner than HSPs, but by less than an order of magnitude; in other words, these two bumps are much closer in height to the HSP bump than to the UP bump. Although there are other kin-types with equal expected relatedness to, say, HCPs, they require repeat matings in different years, so are ignorably rare.

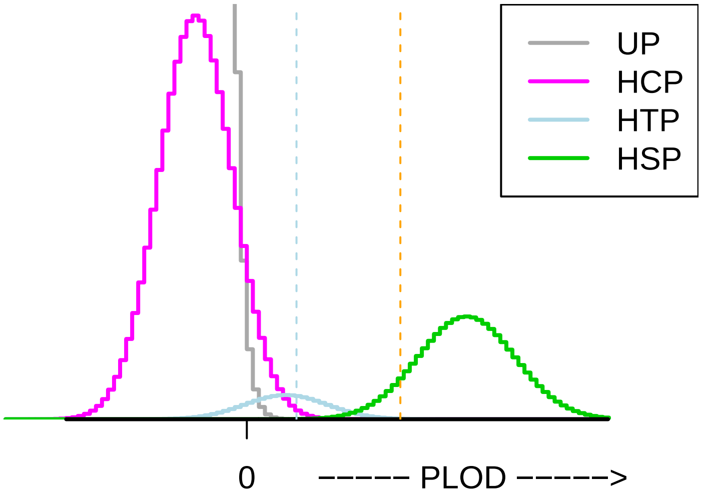
Since the first two cases are straightforward, we concentrate here on the third, less-related, case. There is little prospect of definitively identifying HTPs or HCPs using the approach developed for HSPs. However, the risk is that the HCP or HTP bumps could overlap the threshold η used to exclude false-positives from UPs. Figure 4 shows a possible scenario. The locations of the HCP/HTP bump-centres are completely predictable from theory, but the theory does not *completely* fix their widths or heights. Qualitatively, they will be wider than the UP bump but somewhat narrower than the HSP bump, because the effects of linkage are reduced by crossovers during the extra meioses. The exact widths cannot be predicted accurately even when the overall effect of linkage on the HSP bump is known, i.e. after the width of the HSP bump (HSP PLOD variance) has been estimated in step 4 above. However, with some fairly complicated maths it is possible to use the HSP PLOD variance to bound PLOD variances for HCPs and HTPs (i.e. the bump widths). By combining the upper-bound variance with rough demographic bounds on the relative numbers of HTPs-to-HSPs and HCPs-to-HSPs, one can pick a threshold η' which should exclude almost all false-positive "HSPs" that are really more distant-kin. It is not necessary to choose η' precisely; the important thing is to err on the side of safety and make sure that η' is too high rather than too low. As with UPs, compensating for false-negative HSPs with PLOD $< \max(\eta, \eta')$ is straightforward, and (if enough loci are used) incurs only a mild efficiency cost through loss of potential information from a moderate number of kin.

Kin-types more distant than HCP are essentially irrelevant; while they may be more common than HCPs they will not be greatly so, so their bumps will not be much larger; their bumps are narrower because of diminished linkage effects; any greater numerical abundance is overwhelmed by the greater

[8] "Thiatic" is a close-to-genderless term based on the Greek $\theta\iota\alpha/\theta\iota o\varsigma$ for Aunt/Uncle.

[9] At least when the birth-dates of the pair are fairly close, as they will be at the start of a CKMR study, because the "grandparent" has to survive for at least one extra generation before spawning successfully again.

FIGURE 4. Other types of kin. Blue dashed line shows η for UPs; orange dashed line shows η' for HCPs/HTPs. Here, the main risk of contamination is from HTPs, but the widths and heights are fairly arbitrary (and the UP curve especially would be far larger) so the relative positions of the dashed lines could be quite different in reality.



distance of their bumps from the HSP bump, combined with the superexponential rate of decay in tail probabilities. In effect, distant kin just blend into the UP bump; the only relevant kin-types for false-positive concerns are UPs because there are so many of them, and HCPs/HTPs because they are fairly closely related.

Since the genotyping method and the suite of loci need to be selected *before* any HSPs are available, i.e. before the effects of linkage on HSP PLOD variance can be estimated from data, there is a slight theoretical risk that the HTP or (less likely) HCP bump could overlap with the HSP bump. The critical point would come if η' reaches the HSP PLOD mean; even this should be survivable since a clear bump of HSPs should still be visible, 50% of the true HSPs would still be available, and it would still be possible to estimate the PLOD variance. Happily, such a high η' seems most unlikely. For a tuna with 24 chromosomes, even if there were *no* crossovers (the most conservative assumption, but a biological impossibility), a simple binomial calculation shows that under 1% of HTPs would have a true relatedness as high as the PLOD mean for HSPs. With crossovers at human rates, the overlap between HTP and HSP curves (which is roughly the proportion of false-negative HSPs) might be around 3%. In practice, we cannot know the true relatedness for each pair but instead have to estimate it using a finite set of loci, which will make the bumps somewhat wider and the actual overlap somewhat larger, but certainly there seems no reason for concern about viability. Given the abundance of ABFT, we suspect that UPs rather than HCPs/HTPs will be the dominant source of overlap (i.e. that $\eta > \eta'$).

2.5. **Utility of physical-linkage data.** The ideal basis for deciding on HSPness would be the proportion of the pair's genomes that is coinherited, a.k.a. Identical-By-Descent (IBD). But IBDness cannot be ascertained directly; all we can do is examine "identicality" of alleles, which could happen either by coinheritance or by inheritance from different ancestors who just happened to pass on the same allele. With biallelic SNPs, the latter is a very common event, and as we have seen, it therefore takes a large number of SNPs to distinguish HSPs from UPs.

However, if several SNPs are situated close by on a small fragment of the genome, then either all of them or none of them will be coinherited[10]. If we examine one such fragment with say 20 SNPs in two animals, and find that all 20 do have an identical allele, then it becomes rather likely that the fragment is coinherited. Since a truly unrelated pair will not have *any* coinherited fragments, whereas typically 50% of fragments will be coinherited in HSPs, finding even one coinherited fragment is quite informative— much more so than just checking 20 SNPs scattered across the genome. In principle, with a moderate number of SNPs on a moderate number of fragments, we might be able to do a more precise job of identifying HSPs; this means a more efficient use of the samples, since fewer HSPs would be sacrificed as false-negatives.

There are several ways that such physical-linkage (i.e. fragment) information might be obtained, e.g. in the course of genome-assembly (Yandell and Ence, 2012). However, it is certainly not essential for finding HSPs, and our view has been that developing an assembly, and the methods to use it, is an unnecessary expense and distraction for CKMR— at least with the type of SNP-based GBS genotyping method we are are using.

Nevertheless, if physical-linkage information becomes available later, say via a genome assembly with "high N50" (i.e. big fragments), then it may prove useful for lowering the false-negative rate without needing to change the genotyping, just by using the knowledge that particular sets of loci are likely to be coinherited. If the false-negative rate is already low, though, the efficiency gain would be small.

2.6. **HSP-finding experiences at CSIRO.** This is a brief summary of our experiences at CSIRO, starting in roughly 2013, of HSP-hunting in four species (we have several more under development). The point is not to give exhaustive details of our methods, nor to claim that they are the only possible route to success, but rather to show that the steps in section 2.1.1 can be implemented successfully— and also that it would be entirely possible to fail, via inadequate QC, insufficient loci, and/or insufficiently precise genotyping.

So far we have interim or final results[11] for three shark species and near-final results for SBT (where we have not yet genotyped enough samples to see HSPs). For the sharks, where population sizes are much smaller, we used "DArTseq(TM)", a GBS method that includes a double-digest complexity-reduction step. Broadly similar patterns were found in all species, with good repeatability, null alleles present at detectable frequency though not especially common, usually only one SNP per locus, and overall low SNP density on the genome (for one species, with very small population size and very recent evolutionary history, it was necessary to genotype twice with different sets of restriction-sites in order to get enough

[10] Unless a crossover has occurred, which is unlikely unless the fragment is very long.

[11] I.E. to the point of HSP-identification; the ensuing CKMR models are either in submission or in preparation.

SNPs). After QC, we were left with 1200–1500 satisfactory loci. The initial analysis for two species showed that the UP bumps and HSP bumps were not quite where they should be, but this proved to be a result of large-scale population structure (i.e. different allele frequencies in different subpopulations), and was eliminated on reanalysis of subsets. For the one shark species where we so far have completed all analyses, there was no problem picking thresholds leading to negligible false-positive and low (<5%) false-negative rates; the η'-threshold for HCPs happened to fall just below the η-threshold for UPs.

With SBT, we began genotyping in 2006 using the only proven method then available (microsatellites; Bravington et al. (2016a)). In 2013, we decided to switch to what eventually became "DArTcap(TM)"— even though this would entail the pain and expense of re-genotyping most of our existing samples. There were three or four reasons, including long-term cost-savings (Bravington et al. (2014); Bravington et al. (2015)), but the most important was that a SNP-rich method would be the only way to unleash the power of HSPs.

The large sample size for SBT also dictates a slightly different approach than for sharks, because of both cost-scaling and increased need for genotyping accuracy with a much larger "UP bump". We are now using "DArTcap(TM)", which follows the double-digestion by a capture-probe step to focus on selected loci and give high read-depth for low cost. We selected loci for capture based on a pilot DArTseq run similar to the shark process in 2014/5, which we used to largely eliminate paralogous loci and pick loci with statistically informative Minor- and Null-Allele Frequencies. The SBT work is not yet finished (there are about 20,000 samples to genotype; Farley (2016) and Davies (2016)); the summary here encompasses a pilot DArTcap study in 2015 using about 600 loci and 200 samples (Bravington et al. (2015)), plus a preliminary examination of the larger set of loci that are being sequenced on the full set of samples.

The SBT genome— which is presumably similar to other bluefin genomes— looks rather different to the shark genomes. SNPs seem to be much denser in SBT, and most of our loci have occasional mutations at extra sites as well as the "target" SNP. Null alleles— presumably from mutations at restriction sites— are very common, and show high repeatability and heritability, as confirmed on our set of known POPs from microsatellite analysis. Average read-depth varies across loci, but for most loci the mean read-depth among heterozygotes is several hundred, which permits fairly reliable discrimination between single-null and homozygote genotypes (e.g. AA vs AO). Although this discrimination is not perfectly accurate, the error rate can be estimated and allowed for when constructing the PLOD. Other types of error rates, e.g. in calling heterozygotes or double-nulls, appear to be negligibly low.

Based on the pilot study (Bravington et al. (2015)) we expect to find about 1500–1600 usable loci for SBT, and the UP bump matches the theory; we do not yet have enough genotypes to see the HSP bump. It turns out that these 1500 loci should be ample for HSP-vs-UP identification in SBT (i.e. η will be well to the left of the HSP bump's mode) but *only* because of the ability to discriminate AA from AO pairs; if read-depths were lower so that we were forced to genotype such cases as "either AA or AO" without discriminating, then we would need more loci (and have to pay more) to avoid the situation of Figure 2. Of course, the frequency of null alleles depends on the genotyping method, but with DArTcap we have been able to turn the apparent problem of frequent nulls into a considerable advantage.

So far, we have been pleased with the results from DArTcap, but of course genotyping technology continues to evolve, and better methods may yet appear. Nevertheless, we did not in 2013 (and do not now) expect to *need* to change genotyping methods again, as we eventually had to with microsatellites, at least not for "mainstream" teleost and shark projects. Thanks to having HSPs as well as POPs, the information we can now get for CKMR is sufficient for most assessment and management situations that we have contemplated, and the unit cost of genotyping is acceptably low for most species. Costs are likely to drop within-method as well as between-methods, since sequencing itself continues to become cheaper.

In order to use the genotype data in CKMR, we have also developed in-house R code for the HSP-finding steps, since existing kinship software is not suitable for this specialized task (see A.4). We expect to publish an R package and accompanying paper in late 2017.

3. Outline for a genotyping workshop

By now it should be evident that genotyping-for-HSPs is a reasonably complicated topic. A workshop would be a very useful way to explore the issues for different genotyping methods as they pertain to CKMR. There is at least one candidate method (the DArTcap approach we use at CSIRO), but there might be others too— and of course finding kin-pairs is not the only genetic issue of importance for ABFT. We expect that any useful workshop on genotyping issues for ABFT CKMR would need to address at least the points below. The prerequisites for the workshop should be clear from the list. Of course, before any workshop, this list would evolve through discussion with interested parties.

Objectives for genotyping: Kin identification as discussed in this document would only be part of the story for ABFT CKMR. At least three other objectives of genotyping would need to be considered, perhaps requiring different/additional techniques:

- **mtDNA::** Analysis of mtDNA is essential for HSP-based CKMR, since it serves to distinguish (at least probabilistically) maternal from paternal descent for identified HSPs. Allowing for sex differences is important in avoiding bias in CKMR. It is necessary to decide on a genotyping method (usually, different from the way nuclear DNA for kinship is analysed), on which samples need to be analysed, and on how thoroughly to resolve haplotypes.
- **Stock markers::** Distinguishing EABFT samples from WABFT (at least probabilistically) is essential for CKMR for ABFT. The genotyping method is not necessarily the same as for kin-finding per se. It is necessary to decide whether all samples are to be genotyped (e.g. would there be any point for 1yo in the Med?) and how.
- **Other markers:** e.g. genetic sex markers— useful if available, but how to genotype?

Principles and practice: for HSP identification, and how to assess adequacy of a genotyping method. Presumably this would build on the material in this report. There would be no point in holding a workshop before suitable software exists for doing the calculations in section 2.1.1, and participants would need appreciable time beforehand to prepare performance data.

Performance data: for each proposed genotyping method. Some or all of the data might have to be extrapolated from preliminary work, rather than being based on any "final set" of loci. To assess adequacy for HSP-finding following the statistical principles in section 2, the minimum requirements would be at least:

- expected number of loci
- allele frequencies per locus, including of null alleles
- estimates of genotyping error rates, for different types of error
- costs
- general diagnostic information on locus selection and compliance with theory, such as observed and expected genotype frequencies.

Other issues: relevant to choice of method and feasibility of CKMR implementation: e.g. value or otherwise of genome assembly; curation of data; QC procedures; logistics.

4. Summary and Recommendations

If CKMR could be successfully implemented for EABFT, then it would prove of great benefit to assessment and management; CKMR has the potential to resolve key uncertainties around abundance, stock structure, and fecundity in EABFT which do not seem tractable with any other type of data. Two key issues, though, would need to be addressed before CKMR could be successfully implemented:

- selection of a genotyping method for kin-finding which can find Half-Sibling Pairs with quantifiable reliability;
- detailed design of a sampling scheme (a refinement of Davies et al., 2017), to ensure adequate sample sizes, appropriate ancillary data, quality control, and management of samples.

The statistical principles of finding HSPs with modern genotyping methods, and some of the practical issues that need to be explored within any particular method, are described in section 2 of this document. From our own experience to date, we know that at least one adequate genotyping method does now exist (which would not have been the case 5 years ago). However, there may be other genotyping methods to consider, and the technical issues are complex. We suggest that the best way to start the decision process would be with a small technical workshop to explore the issues in more detail, and to compare preliminary results from any proposed methods (section 3). Considerable preparation time would be needed to get preliminary results for any method that is to be considered at a workshop (and, in particular, no suitable software is generally available yet). Perhaps an achievable timeframe would be in late 2017 to confirm interested participants, methods, and preparatory activities, with the workshop itself scheduled for mid 2018 at the earliest. We note that there is growing interest worldwide in CKMR for tuna and beyond, and hence that there may be parallel initiatives which could benefit from a combined effort.

The second issue, of detailed sampling design beyond the broad outline of Davies et al. (2017), is largely beyond the remit of this document. However, it is worth here touching on implications for genotyping and timelines. The overall sample sizes required for CKMR EABFT suggested in Davies et al. (2017) are modest given the value of EABFT fisheries and the ongoing uncertainty in key assessment parameters. The qualitative data requirements of CKMR are also not much beyond what is needed for assessing most modern industrial fisheries (CKMR needs good-quality tissue samples, in addition to the usual length measurements and otoliths, and preferably total catch but no need for effort). Although this does not seem overwhelming, there are substantial complexities around EABFT breeding biology and fishery sampling in the Mediterranean, leaving an open question of whether a comprehensive CKMR program in the Mediterranean is currently possible. A staged implementation might be useful— it could help iron out to the sampling and genetic and statistical issues, and to answer some of the basic biology so that any future designs can be streamlined— but even then would need careful design. While there is considerable flexibility in CKMR design (much more so than in, say, line-transect surveys), there are limits beyond which no useful information will be returned. The stock structure discussion in Davies et al. (2017) shows how a parameter like SSB can become inestimable if some key types of CKMR sample are missing, regardless of how well sampling is conducted elsewhere; what constitutes a "key type of sample" can, of course, only be explored through design. In terms of genotyping and timelines, the salient points are that:

(1) CKMR sampling design, and at least the early stages of sample collection, are largely independent of what genotyping technique subsequently gets used. (A possible exception could be if especially delicate samples are proposed, e.g. larvae; such complications are probably best avoided if alternatives can be found.)

(2) There is no immediate urgency to make a final decision about genotyping method, because of the time that will be required to do further design and then to accumulate useful numbers of samples. There should thus be ample time to do the preparation required for a successful workshop, to do any further testing required, and to make a considered decision on a genotyping method that will deliver the necessary genetic precision at reasonable cost.

Our recommendation for a way forward on genotyping is as follows:

- Convene a technical committee to develop ToR and activities for a work program leading up to a technical workshop provisionally in mid 2018, aimed at selecting an appropriate genotyping method for EABFT CKMR, as well as methods for other genotyping requirements such as E/W stock markers. The technical committee would need to: seek participants willing to explore specific genotyping methods; oversee preparation of inputs/analysis along the lines of section 3; organize the agenda and the workshop itself; and keep track of any similar initiatives internationally, since there is no point in duplicating efforts.

5. Acknowledgements

Pierre Feutry is thanked for helpful comments on an earlier draft of this report.

Appendix A. Background

A.1. **PLODs.** In the parlance of statistical genetics, the "LOD score" for one locus ℓ in a pairwise comparison between animals i and j is the log-likelihood-ratio:

$$\text{LOD}_{\text{HU}\ell ij} = \log \frac{\mathbb{P}\left[g_{i\ell}, g_{j\ell} | K_{ij} = \text{HSP}\right]}{\mathbb{P}\left[g_{i\ell}, g_{j\ell} | K_{ij} = \text{UP}\right]} \tag{1}$$

where $g_{i\ell}$ and $g_{j\ell}$ are the *observed* genotypes for i and j at locus ℓ, and K_{ij} is the true (but unknown) kinship of i and j. The probabilities on the numerator and denominator can be computed from basic genetic and statistical principles— remembering to allow for genotyping error probabilities, if substantial— and assuming that allele frequencies are known accurately, which will be the case whenever reasonably large samples are collected.

Note that the definition of "the" LOD depends on the two hypothesized kinships in the numerator and denominator, in this case HSP and UP; if we were instead trying to distinguish between, say, an Aunt-Nephew pair and an Unrelated pair, then we would in principle compute a different LOD. Since this document is mainly concerned with HSP-vs-UP comparisons, we generally refer just to "the" LOD or "the" PLOD and omit the subscript "HU"; in practice, the LOD for HSPs is quite effective for distinguishing between several types of close kin. Note also that the true kinship K_{ij} may not be one of the two possibilities used in computing the LOD (section 2.4).

A positive LOD is evidence in favour of HSPness (or at least some degree of kinship), and a negative LOD is evidence in favour of UPness. In fact, the LOD is statistically optimal for choosing between those two options. Of course, with just one locus, the statistical power to discriminate between HSP and UP would be negligible, so it is necessary to somehow combine LODs across loci. If the loci were statistically independent (i.e. if the pair of genotypes at different loci were statistically independent), then the sum of the LODs across all loci for each pair of animals would itself constitute a LOD; but independence only applies when (i) the pair is Unrelated so that linkage (A.2) does not apply, and when (ii) none of the loci are in Linkage Disequilibrium (A.3). Regardless of independence, the sum of the per-locus LODs is nevertheless a good test statistic for HSPness vs UPness, with known distribution under the null hypothesis (denominator) of UPness, and known expectation under the alternative hypothesis (numerator) of HSPness; in fact we suspect sum-of-LODs is optimal, since it is difficult to see how it might be improved in the absence of detailed linkage information. However, since the sum is not generally a true LOD, we refer to the sum-of-LODs as a PLOD (Pseudo-LOD):

$$\text{PLOD}_{ij} = \sum_{\ell} \text{LOD}_{\text{HU}\ell ij} \tag{2}$$

A PLOD value of zero corresponds to pairs-of-genotypes that have the same overall probability-of-occurrence (across all loci) for a UP as for a HSP. That is *not* the same thing as saying that a pair of animals with an observed PLOD of 0 is equally likely to be UP as HSP, because the prior probability of UPness is vastly greater (i.e. the bump on the left is always much bigger than the bump on the right).

A.2. **Linkage.** DNA is organized into a moderate number of continuous strands— chromosomes— and most cells contain two different copies of each chromosome. Inheritance involves meiosis, whereby each egg or sperm cell acquires a *single* copy of the parental DNA, i.e. just one version of each chromosome. If there was no crossover (explained soon) during meiosis, then the HSP would co-inherit either all of, or none of, the loci on each chromosome— so in a species with only say 4 chromosomes, each HSP would co-inherit a proportion of exactly 0/4, 1/2, 2/2, 3/3, or 4/4 ... of their loci. With so few chromosomes, there would be an appreciable chance (6%) that all loci would co-inherit, in which cases that particular HSP would be just as genetically related as a Parent-Offspring Pair— and a similar chance that an HS pair has *no* coinherited genetic material. The "genetic relatedness" of HSPs is thus quite variable, even though the biological kinship is always the same. For this reason, it is impossible ever to be completely certain that *all* HSPs have been found, no matter how thorough the genotyping is. This invites the question of whether to try making definite categorizations of kinship at all; perhaps one could instead work directly with the observed genotypes? That turns out to be a bad idea; see Bravington et al. (2016b). Instead, the key is to take an engineering perspective: go ahead with definite categorizations but acknowledge that they might be wrong sometimes, and design the whole process to have a quantifiable and acceptably low overall error-rate in HSP identification.

In practice, the variability-in-relatedness of HSPs and other kin is mitigated since most species have many more than 4 chromosomes, and there are often one or more "crossovers" per chromosome per inheritance (i.e. per meiosis), whereby the inherited part switches from one copy to the other at some random point along the chromosome. Humans have 23 chromosomes, with on average 1–2 crossovers per chromosome per inheritance; *Thunnus* spp are reported to have 24 chromosomes (Ida et al., 1991; Soares et al., 2013), with unknown crossover rates.

"Linkage" can be seen as the conditional probability that one locus will be coinherited, given that another one is. Loci on different chromosomes are completely unlinked, but loci on the same chromosome will be somewhat linked, the extent depending on the phyiscal proximity of the loci and the crossover rate. In the context of HSP-finding, it is essential to be *qualitatively* aware of the existence of linkage and to allow for it statistically, but it is *not* necessary to estimate actual linkage rates between specific loci— that is a colossally difficult exercise of great importance in tomato-breeding etc., but not in stock assessment.

A.3. **Linkage Disequilibrium.** A new SNP allele that arises from mutation will occur on just one strand of parental DNA, and when that strand is inherited, the new allele will be linked with one particular allele in all nearby loci on the same strand. Consequently, the presence of the new SNP is statistically predictable given the value of neighbouring loci. This linkage will persist for several generations, but eventually, accumulated crossovers will break up the linkage so that the presence of the new SNP is no longer predictable based on distant loci. For close-by loci, though, the chance of intervening crossover breaking the linkage is so small that predictability will remain for many generations even in a large population. The allele distribution at the new SNP site is not statistically independent of the alleles at neighbouring loci, and this is Linkage Disequilibrium (LD; Slatkin, 2008); it applies at a population level, whereas linkage *per se* applies only to kin. LD among the loci used for genotyping would reduce statistical power for kin-finding, because it would increase the PLOD variance. The genetic distance required for LD to disappear between loci depends on the species, but should not pose any problem for just a few thousand loci in an abundant freely-breeding species with substantial evolutionary histories, such as tuna.

A.4. **Software.** Kin-finding is of widespread interest beyond CKMR, and there is existing software (ML-RELATE, COLONY, CERVUS, etc). However, no existing program that we are aware of seems suitable for finding HSPs for CKMR, for several reasons:

(1) Exceedingly tight control of false-positive rates: the enormous size of the UP bump requires the ability to control false-positives down to about the 1-in-100,000,000 level, and it is also essential for CKMR to be able to estimate reliably how many HSPs might be missed due to false-negative "errors"[12]. Existing software tends instead to focus on maximum-likelihood assignment of specific pairs, anticipating false-positive/negative rates of the order of 1 in 100.

(2) Huge number of pairwise comparisons: for EABFT CKMR, over 100,000,000 pairwise comparisons would be needed to find a statistically useful number of HSPs (we used 38,000,000 comparisons for SBT). With 1000s of loci considered for *each* comparison, speed is a serious issue.

(3) Limited targets: the only kinship of real interest is HSP, so an efficient algorithm can be designed specifically for that case. Other kinships are either obvious (POPs, FSPs) or not of direct interest (HCPs etc) and can be excluded *post hoc* by careful thresholding (section 2.4).

(4) Lack of family groups: with large populations like tuna, it is a rare event even to find a pair of kin, and triads etc are negligibly rare (of the order of 1/population-size, as common as HSPs). For CKMR, there is no point— and no chance computationally— of addressing the huge complexities of family-reconstruction that is attempted by full-pedigree methods like COLONY.

(5) We need the ability not just to analyse data, but also to predict in advance the effectiveness of a suite of loci / a genotyping method for HSP.

For our work on sharks and SBT at CSIRO, we have developed R code which addresses the above points and implements steps 1–7 of section 2.1.1. The inputs are sample genotypes and error-rate estimates; genotyping per se, and error rate estimation, have to be done separately. The code currently handles biallelic SNPs with allowance for nulls, so in principle it is applicable to many but not all GBS methods. We expect to release the R package and accompanying paper in late 2017.

References

Bravington MV, PM Grewe, and CR Davies (2016a). "Absolute abundance of southern bluefin tuna estimated by close-kin mark-recapture". *Nature Communications, 7, 13162.*

Bravington MV, HJ Skaug, and EC Anderson (2016b). "Close-Kin Mark-Recapture". *Statistical Science.*

[12] They are not really "errors", since they are anticipated and allowed for in the steps of section 2.1.1.

Bravington M, J Eveson, P Grewe, and C Davies (2015). *SBT Close-Kin Mark-Recapture: options for the medium term.* Tech. rep. CCSBT-ESC/1509/19. Commission for the Conservation of Southern Bluefin Tuna.

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

Davies CR (2016). *Next-generation close-kin mark recapture: using SNPs to identify half- sibling pairs in Southern Bluefin Tuna and estimate abundance, mortality and selectivity. FRDC/CSIRO project (2016-044).* Tech. rep.

Davies C, M Bravington, and R Thomson (2017). *Advice on Close-Kin Mark-Recapture for estimating abundance of Eastern Atlantic Bluefin tuna: a scoping study. [Revised].* ATLANTIC-WIDE RESEARCH PROGRAMME ON BLUEFIN TUNA (ICCAT GBYP– PHASE 5) GBYP 07c/2015. International Commission for the Conservation of Atlantic Tuna.

Ellis JS, J Gilbey, A Armstrong, T Balstad, E Cauwelier, C Cherbonnel, S Consuegra, J Coughlan, TF Cross, W Crozier, et al. (2011). "Microsatellite standardization and evaluation of genotyping error in a large multi-partner research programme for conservation of Atlantic salmon (*Salmo salar* L.)" *Genetica* 139.3, pp. 353–367.

Farley J (2016). *Annual close-kin tissue sampling, processing and Indonesian ageing project. CCSBT-CSIRO project for monitoring spawning stock.* Tech. rep.

Ida H, N Oka, and Ki Hayashigaki (1991). "Karyotypes and cellular DNA contents of three species of the subfamily *Clupeinae*". *Japanese Journal of Ichthyology* 38.3, pp. 289–294.

Slatkin M (2008). "Linkage disequilibrium— understanding the evolutionary past and mapping the medical future". *Nature Reviews Genetics* 9.6, pp. 477–485.

Soares RX, LAC Bertollo, GWWF da Costa, and WF Molina (2013). "Karyotype stasis in four Atlantic *Scombridae* fishes: mapping of classic and dual-color FISH markers on chromosomes". *Fisheries Science* 79.2, pp. 177–183.

Thompson EA (2000). "Statistical Inference from Genetic Data on Pedigrees". *NSF-CBMS Regional Conference Series in Probability and Statistics* 6, pp. i–169.

Yandell M and D Ence (2012). "A beginner's guide to eukaryotic genome annotation". *Nature Reviews Genetics* 13.5, pp. 329–342.