OCEANS AND ATMOSPHERE
www.csiro.au

SBT kin-finding and genotyping update

Mark Bravington

CCSBT-OMMP/1706/??

Prepared for the 8th Operating Model and Management Procedure Technical Meeting

held in Seattle, U.S.A. 19th to the 23rd of June 2017

CSIRO
Oceans & Atmosphere
Castray Esplanade, Battery Point, Hobart TAS 7000, Australia
Telephone : +61 3 6232 5222

Fax : +61 3 6232 5000

SBT KIN-FINDING AND GENOTYPING UPDATE

MARK BRAVINGTON, JUNE 2017

Abstract. After a lot of work to change genotyping methods and develop suitable algorithms, we have succesfully found POPs and HSPs with SBT samples from 2006 thru to 2015, and are ready to incorporate them into CKMR models. See Figures 2.1 and 3.1 for visual confirmation.

Contents

1. Introduction

CKMR uses modern genetics to find pairs of close relatives amongst large collections of tissue samples. It then embeds the number of such pairs found, along with covariates such as their age and sex and date of collection, in an extended mark-recapture framework, where "recaptures" are kin of an animal rather than the animal itself. This allows direct estimation of adult abundance and other demographic parameters without needing to rely on CPUE or catch data. Bravington et al. (2016b) provide an overview and technical details.

CKMR has proved very useful to CCSBT for providing fishery-independent monitoring of spawning-stock biomass, and for establishing key biological parameters relevant to management, such as fecundity-at-age. The initial application to SBT, based on Parent-Offspring Pairs (POPs) identified using microsatellite loci amongst juveniles collected from the GAB and adults from Indonesia between 2006 and 2010, is described in Bravington et al. (2012), Hillary et al. (2013), Bravington et al. (2014), and Bravington et al. (2016a). Sample collection has continued thereafter, in anticipation that more CKMR analysis would be needed at some point in the future, but no further genotyping was done for a few years.

In 2014 and 2015, CSIRO proposed changing the genotyping technique in future from microsatellites to a modern Genotyping-By-Sequencing (GBS) method now known as "Dartcap" (TM), which uses SNP loci rather than microsatellites; see Bravington et al. (2015) for more details. There were three independently compelling reasons:

(1) future-proofed
(2) cheaper
(3) able to find Half-Sibling Pairs (HSPs) as well as POPs

Management-wise, the last point is the most important, since having HSPs will permit a direct estimate of adult mortality rate, without requiring untestable assumptions. Thus, selectivity on adults can in principle be separated from natural (and fishing) mortality, which is a notoriously tough problem for fisheries in general, let alone for SBT.

This proposed change-of-genotyping-method was agreed, and CSIRO eventually secured the funding to go ahead with the requisite genotyping in time for the 2017 CCSBT stock assessment. Specifically, this covered:

(1) juveniles in the original 2006–2010 samples (since their parents might still be found in post-2010 adult samples);
(2) the "back-catalogue" of as-yet-ungenotyped samples from between 2011 and 2015.

1.1. **Terminology and timeline so far.** The "genotyping" process can be split into several parts. The first part, here called "sequencing" for brevity, is carried out by Dart Pty Ltd (Canberra), who have developed the specific variant of GBS that we are using. The second part, "calling", turns the data-files from sequencing into a *genotype* for each sample and locus, ie the pair of alleles inferred to be present; this entails some quite complicated algorithms developed at CSIRO specifically for Dartcap sequencing data. A byproduct is estimates of the genotyping error rates for each locus. The final step prior to CKMR itself is kin-finding, based on the inferred genotypes and the error-rates. For that, we have developed generic algorithms (i.e. not specific to Dartcap) from basic statistical principles, which are summarized below and/or in section 5 of Bravington et al. (2016b). Control of false-positive and false-negative rates is *crucial* to kin-finding, since ~100,000,000 comparisons might be needed to find ~100 kin-pairs; the vast majority of comparisons will be UPs (Unrelated Pairs)

Details of "sequencing", including choice of loci, were tuned during 2016; after checking preliminary results, and finalizing the funding arrangements, large-scale sequencing of around 16,000 fish began in February 2017; and the full set of sequencing-files were received by CSIRO at the end of March 2017. In parallel, CSIRO was developing QC, "calling", and kin-finding algorithms suitable for this newest type of sequencing data. From April until now, we have been refining and applying those algorithms to the new data. We now have reliable sets of POPs and HSPs suitable for CCSBT use; the POPs have already been used in OM updates (Hillary et al., 2016), but there has not yet been time to incorporate the HSPs. See Section 4 for further work.

Table 1 summarizes the available samples from Dartcap, excluding about 700 that have been rejected so far on QC grounds. (There are also about 5,000 replicate genotypes, which can be used for estimating error rates.) For the sake of economy, we did not re-genotype adults from 2006–2009, since any usable offspring would already have been found in the original microsatellite genotyping[1] (except that 2009 adults could still have undetected 3yo offspring caught in 2011). For CKMR models, therefore, the new POPs and HSPs have to be combined with the old POPs from 2006–2010 samples.

2. POP results

The microsatellites used in the first round of SBT CKMR were adequate for finding POPs using Mendelian-exclusion principles; see long appendices in Bravington et al., 2014. However, a lot of statistical processing and lengthy explanations were entailed to control false-positive rates and demonstrate that false-negatives must be rare; in short, we did have enough loci to do it, but only just. The DartCap genotyping has been designed with the goal of identifying HSPs, which is much harder than finding POPs; consequently, finding POPs ought to be easier and clearer now. While this is broadly true, there are some details arising from the switch to SNPs which require attention. I first show the results, and then briefly explain the modified exclusion criterion developed for Dartcap.

[1]We deliberately do not check for POPs if the adult was caught in the season the juvenile was spawned; see previous CCSBT documents for reasons.

TABLE 1. Number of samples genotyped by year and origin (after some QC checks)

	Indo	Port L
2006	0	1281
2007	0	1305
2008	0	1315
2009	0	1317
2010	943	1284
2011	931	938
2012	527	844
2013	933	873
2014	904	873
2015	0	922
Total	4328	10952

Figure 2.1 shows part of the histogram of modified-exclusion-statistic across all Dartcapped adult-juvenile pairs. The POPs are visible as a small bump on the LHS. It is impossible to plot the entire histogram on one figure, because it is entirely dominated by the gigantic bump of UPs whose peak is around 0.116 (exactly where theory predicts it should be, based on allele frequencies of each locus). The giant bump drops off very quickly, and the flattish tail around 0.055–0.075 will contain a number of adult/juve HSPs or GGPs (Grandparent-Grandoffspring Pairs), which should be somewhat rarer than true POPs on demographic grounds. Note that the exclusion criterion is not designed to detect HSPs, and is not particularly sensitive since it only uses "simple" 4-way classification of genotypes (see A), so it is encouraging to see that HSPs are almost evident even with the POP criterion. The real point of the figure is that POPs are *clearly* separated from non-POPs— this is much more obvious with Dartcap data than it was with our microsatellite data. The 1500 low-information SNP loci from DartCap are doing better than 25 high-information microsatellite loci, at about half the cost.

As per Table 1, this plot excludes the POPs already found via microsatellites. However, as a check we also Dart-capped those particular pairs-of-samples, and all of them clearly came up as POPs this time too. Interestingly, at least to us: we also Dart-capped one curious adult/juve pair that was clearly not a POP according to microsatellites but nevertheless remarkably close (just two unambiguous Mendelian exclusions in 25 loci compared). And the DartCap WPSEX for this pair was around 0.06, consistent with being a GGP or HSP.

The POPs found this time appear generally consistent with previous results; see Hillary et al. (2016) from which Table 2 is taken. As previously, all identified parents were 8yo or more in the year their identified offspring were spawned.

2.1. **Modified exclusion criterion (technical).** Offspring inherit one allele from each parent at each locus; thus, a POP should share an allele at every locus. "Mendelian exclusion" uses this to test for POPs; in principle, failing to share an allele at any locus (here 1500 of them) means that a pair cannot be a POP. In practice, this needs softening a bit, because of null alleles, the possibility of genotyping errors and (with 1500 loci) even perhaps a mutation or two, all of which could lead to *apparent* exclusions even in genuine POPs.

The exclusion criterion I've used for the SNP loci in Dartcap is slightly different to that used for microsatellite loci in our original study, because SNP loci have only two "normal" alleles (alwass called A & B) and because null-alleles (called O) are very common in SBT. To explain further, section A provides background. Although a key part of DartCap 's lustre is the ability to resolve each genotype into the 6 possible cases AB/AA/AO/BB/BO/OO, for POP-finding I've used a simpler 4-way classification where AA/AO are merged (called "AAO") and similarly for BB/BO. Analysis of replicates shows that error

FIGURE 2.1. POPs via weighted-pseudo-exclusion (wpsex). Low values indicate POPs. X-axis truncated to omit gigantic peak of UPs.

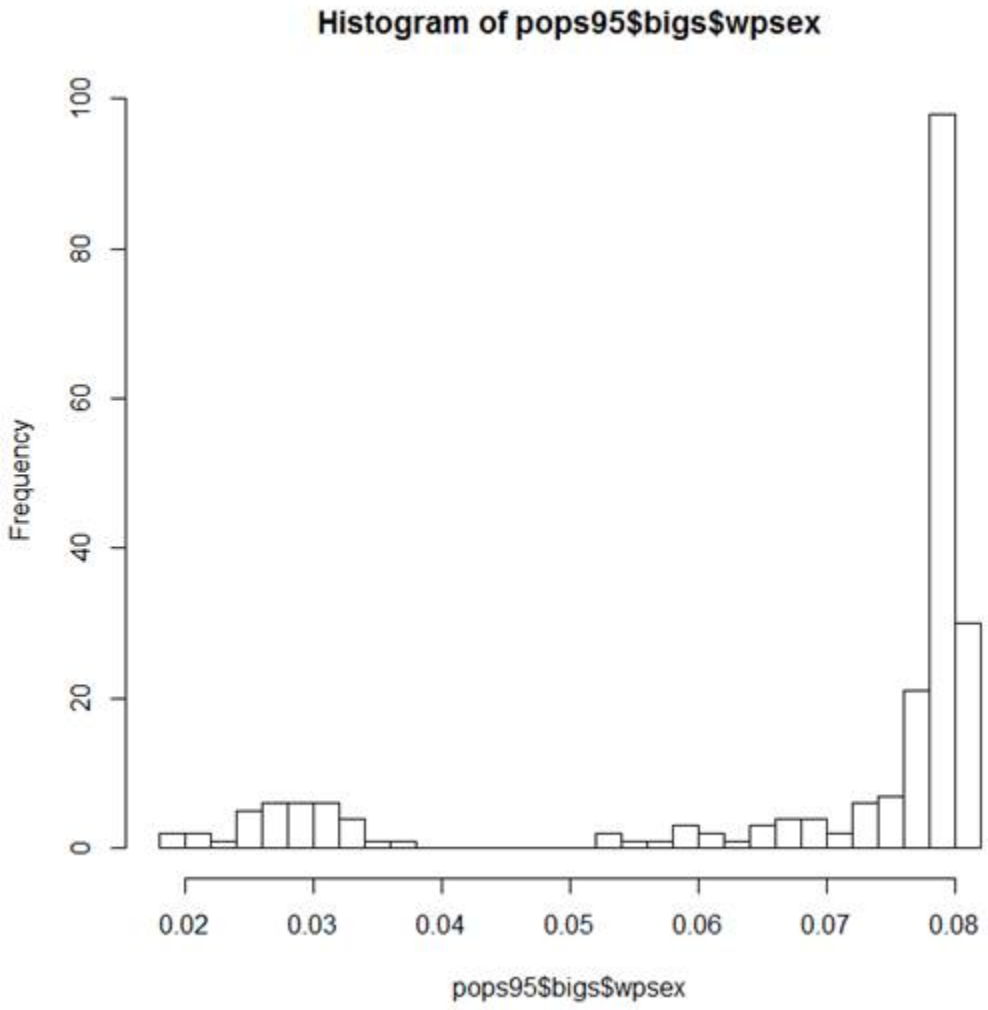


TABLE 2. POP breakdown (new + old). Row = Juvenile birth-year; column = adult capture-year. From OMMP/1706/4.

	2006	2007	2008	2009	2010	2011	2012	2013	2014
2002	0	0	0	0	0	0	0	0	0
2003	0	5	1	2	0	0	0	1	0
2004	0	2	0	0	3	0	0	0	0
2005	1	4	5	4	1	0	1	1	2
2006	0	4	3	2	0	0	0	0	0
2007	0	0	3	4	1	3	2	0	2
2008	0	0	0	0	0	1	1	1	0
2009	0	0	0	0	0	1	1	1	0
2010	0	0	0	0	0	3	1	4	0
2011	0	0	0	0	0	0	1	2	1
2011	0	0	0	0	0	0	0	1	1

rates within these 4 categories are very low— about 0.1% for most loci— and not having to worry about errors simplifies the calculations markedly.

If a locus has very few nulls, then finding that one fish is AAO means it is probably AA. If another fish is BBO, then it is likewise probably BB; so those two fish probably do have an exclusion at that locus. This is not definitive, because the first fish *could* be AO and the second *could* be BO and the O might then be co-inherited. Nevertheless, adding up the number of these "pseudo-exclusions" is an intuitively powerful way to separate POPs from UPs. It turns out that one refinement is necessary; pseudo-exclusions are more informative at some loci than others (depending on the frequency of nulls at each locus, and to some extent also on the frequency of A vs B alleles), and a weighted version is statistically much more powerful. In other words, an AAO/BBO pseudo-exclusion at locus ℓ receives weight w_ℓ (or 0 if no pseudo-exclusion), and for any pair of fish the Weighted-PSeudo-EXclusion (WPSEX) statistic is the sum of all w_ℓ. The weights are chosen to minimize the false-positive probability.

If there were more than two alleles, then other types of definite exclusion could be considered (e.g. AB vs CCO or AB vs CD), which is how we handled the microsatellite data— but this is by definition not an option for biallelic SNP data.

It is also possible to find exclusions where one fish is AB and the other is OO. These are comparatively rare because we deliberately avoided loci with very high null-allele frequency, so on its own the number of AB-OO is not a good single criterion for POP-finding. Nevertheless, the results from n-AB-OO do nicely (and independently) back up the results from the preferred WPSEX statistic, in that POPs have many fewer n-AB-OOs (not shown). Although genotyping errors are rare overall, there are enough to generate a few AB-OO exclusions even among true POPs.

Many authors (including Bravington et al., 2016b) propose likelihood-based criterion for POP-finding, instead of exclusion. In principle, a likelihood-based criterion is more powerful (optimal, in fact) because it uses more information, even for loci where no exclusion is present; if, say, allele B is rare at one locus, then finding B in both animals of a pair increases the evidence in favour of their kinship. However, I have preferred to stick with an exclusion-based criterion for SBT, for several reasons:

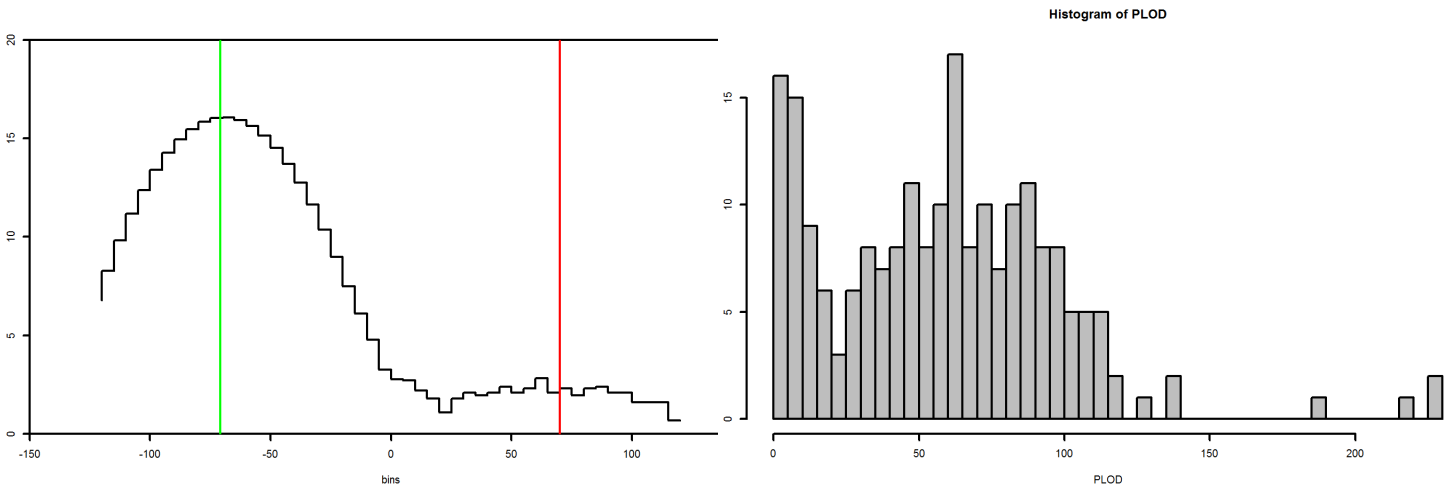
- simplicity;
- plenty of loci, so no need to produce an optimal method (unlike for HSPs, where every bit of data helps);
- no need to rely on estimated rates of *rare* errors.

3. HSP results

Finding HSPs is more difficult than finding POPs, because the degree of kinship is weaker. In general, it is not possible to expect truly 100% reliable ID of HSPs, because the degree-of-relatedness varies randomly between different HSPs, and some pairs may be chance only be weakly related— so, no matter how thorough the genotyping, some overlap with UPs or weaker kin such as first-half-cousin-pairs (one shared grandparent) may be unavoidable. What is important is to be able to set a cutoff for whatever statistic is used for HSPs, so that false-positives from UPs and from other weaker relatives are statistically negligible. Then, provided that the cutoff is well below the mean value expected for a true HSP, it is possible to allow for false-negatives in an unbiased way. For details, see Bravington et al. (2016b) section 5, and Bravington et al. (2017).

In 2015 and 2016, we successfully implemented the approach at CSIRO for several shark species with very low abundances and a more basic version of GBS genotyping. However, to get things to work for SBT where ~100,000,000 juvenile-juvenile comparisons will eventually be made and false-positives are thus much more of a potential risk, it is necessary to have high-precision genotyping (the 6-way null-scoring classification in section A) and a "sensitive" definition of the statistic (the PLOD, or Pseudo Log-ODs ratio) that allows for genotyping error. It also turns out that low-quality samples (contaminated and/or degraded DNA) can become a real problem for spurious HSPs with huge datasets, so extra care

FIGURE 3.1. HSPs. Left: **log** histogram to show all PLOD. Green and red lines are theoretical means for UPs and HSPs respectively. Right: actual histogram of PLODs above zero.



has been needed to filter them out. We are still refining that process, but have already managed to eliminated the problem in practical terms. To cut to the chase, Figure 3.1 presents the PLOD statistic across all pairwise comparisons of 10,809 juveniles. Bigger PLOD values mean more relatedness; the HSP bump on the right is clearly visible, and pretty well separated from the morass of unrelateds on the left.

The four pairs on the far right of the RH figure are presumably Full-Sibs, though I have not checked this yet; in each case, both animals were caught in the same year and are thus likely to be from the same cohort.

The theoretical and empirical distributions of PLODs match very well, though there is still a bump around PLOD~=0 which comes from remaining lower-quality samples and/or from weaker kin. At any rate, that bump has clearly fallen away by three-quarters of the way to PLOD=50, at say PLOD=37. This might be used as the cutoff for false-positives; estimating the false-negative probability is a simple task (by eye, around 15%), but not yet done. Note that the false-negative rate is not very important to inference for SBT, because the *number* of HSPs will likely not be used directly for abundance anyway (see Bravington et al., 2015 for explanation).

Using 37 as a cut-off PLOD for true HSPs, the following summary applies:

- 140 HSPs (possibly including one or two FSPs around PLOD=140, not yet checked).
- All pairs are distinct, except for 3 triplets. In two of those triplets, all the three animals are HSPs, i.e. they share the same parent; For one triplet only, all three animals were caught in the same year. The third triplet very clearly consists just of two pairs A-B and A-C (i.e. B and C are definitely not HSPs), so that e.g. A and B share a Mother, but A and C share a Father. Note that a couple of "accidental triads" are to be expected. About 280 of 10,000 juveniles are involved in HSPs, i.e. 2.8% of animals are in a pair; so, in about 2.8% of the 140 pairs found, the 2nd fish is actually likely to be from one of the other pairs— i.e. will form part of a triad.
- The pairwise tabulation of years (Table 3) shows no obvious pattern. Taking into account the number of comparisons (Table 4), it does look like there is a higher rate for same-year (i.e. same-cohort, mostly) HSPs than for cross-cohort HSPs, suggesting a *small* litter-effect whereby larvae from the same spawning-event may sometimes continue to associate through to the point of capture 3 years later. Overall, though, the proportion of juveniles involved is very small. This is compatible with the usat results, which ruled out a *strong* litter-effect (i.e. it could not be big

TABLE 3. HSPs by years

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2006	2	4	4	4	.	2		.	1	2
2007	.	6	3	6	2	2	2	.	.	2
2008	.	.	4	3	3	3	.	5	1	1
2009	.	.	.	8	6	1	3	7	4	.
2010	.	.	.	.	3	5	3	3	1	3
2011	.	.	.	.	.	6	1	1	2	3
2012	.	.	.	.	.	.	1	2	.	.
2013	.	.	.	.	.	.	.	2	1	2
2014	.	.	.	.	.	.	.	.	3	3
2015	.	.	.	.	.	.	.	.	.	3

TABLE 4. HSP rate per 10^7 comparisons (rounded, and based on small numbers)

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2006	25	24	25	24	0	17	10	0	9	17
2007		72	18	36	12	17	19	0	0	17
2008			49	18	18	25	0	45	9	9
2009				94	36	8	28	62	36	0
2010					37	42	29	27	9	26
2011						141	13	12	25	35
2012							30	28	0	0
2013								53	13	25
2014									81	38
2015										72

enough to have much impact on POP variance) but were not sensitive enough to estimate a low rate directly.

Table 4 is the raw material for estimating adult mortality, based on changing rates of HSP-finding as the gap-in-years increases. However, it would be utterly wrong to do this "by eye" or even by spreadsheet; what the gaps actually show is not the mortality rate, but rather the *turnover rate* in SSB, since if Your parent does not die, it will grow instead and thereby become more fecund and more likely to generate a Sibling for You. While this can be accounted for conceptually using the information from POPs on fecundity-at-age and the length compositions, there is no reliable way to guess the results. The only thing to do is to build and fit a proper CKMR model.

4. Summary and Discussion

Generating and getting to grips with a huge sample of a brand-new type of data— in some respects new even to Dart Pty Ltd— has been challenging and has taken some time. Happily, the exercise has worked; there are useful numbers of POPs and HSPs, and the diagnostics look fine (not all are reported here). There is strong evidence *against* significant overdispersion from "litter effects" of later-in-life sibling association. The POP and HSP data are now ready to use in a stand-alone CKMR model that extends the 2012/2013 POP-only model, as well as in SBT OM work (HSP numbers may change very slightly, but not much). The basic structure of the stand-alone and OM variants needs little change, but a few steps (both in the lab and in software) need completion.

(1) Completion of otolith-ageing for the adults in POPs.

(2) mtDNA analysis of animals in identified HSPs. This determines whether the shared parent is Mother or Father. Male and female SBT exhibit different growth patterns as adults, and may very well have different changes in lifetime fecundity too. This affects the expected numbers and patterns of HSPs, so it is important to analyse separately (e.g. in a stand-alone CKMR model), even if the results subsequently get single-sexed for OM use (as has been necessary with POPs to date).
(3) Selection of PLOD cutoff and estimation of false-negative probability for HSPs.
(4) Adding the log-likelihood for HSPs is pretty simple, in fact arguably easier than for POPs. The basic idea is given in Bravington et al. (2016b), section 3.2.
  (a) Decoupling the selectivity and fecundity terms in the CKMR setup. Some experimentation with functional forms will be needed.
(5) Consideration of whether to allow for ageing-error in juveniles. In the original CKMR study, the length-frequencies suggested that almost all juvenile samples were 3yo. However, since then there have been clear length-frequency shifts in some years, so that an appreciable proportion of the juveniles may actually be 4yo rather than 3yo. This can be addressed if necessary, but is rather tedious.
(6) Allowance for selectivity changes in the Indonesian (in the 2012/2013 CKMR models, Indonesian selectivity was assumed constant on average between 2006 and 2010, but with overdispersion. This is not tenable for more recent years.)

Acknowledgements

Thanks are due to many who made this work possible.

Bali, Indonesia: The Research Institute for Tuna Fisheries (Bali) for collaboration on sampling adults from spawning ground;

Port Lincoln, Australia: the late Tony Jones (Protec Marine P/L) and Adam Kemp at Seatec P/L, for sampling of juveniles; and the Australian Southern Bluefin Tuna Industry Association (ASBTIA) for ongoing support;

Canberra: CCSBT and FRDC for funding;

CSIRO: too many to thank individually: most of all Pete Grewe, Peta Hill, Matt Landsdell for genotyping and logistics, Campbell Davies for oversight, and Paige Eveson for help with developing and testing algorithms.

References

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

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

Bravington, M, P Grewe, and C Davies (2012). *Report of the Close-Kin Project: estimating the absolute spawning stock size of SBT using genetics*. Scientific Committee Report CCSBT-SC/1208/19. Commission for the Conservation of Southern Bluefin Tuna.

Bravington, M, R Thomson, and C Davies (2017). *Genotyping issues for CKMR on Atlantic Bluefin Tuna*. Atlantic-wide research programme on Bluefin Tuna (ICCAT GBYP– PHASE 5) GBYP xxx/2016. International Commission for the Conservation of Atlantic Tuna.

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.

Hillary, R, A Preece, and C Davies (2013). *Updates to the CCSBT operating model including new data sources, data weighting and re-sampling of the grid*. Tech. rep. CCSBT-ESC/1309/15. Commission for the Conservation of Southern Bluefin Tuna.

Hillary, R, A Preece, and C Davies (2016). *Updates required for new data sources and reconditioning of the CCSBT OM*. Tech. rep. CCSBT-OMMP/1706/4. Commission for the Conservation of Southern Bluefin Tuna.

Appendix A. Genetic summary

Bravington et al. (2015) suggested that 1500 loci should be adequate for HSP-finding in HSPs. The sequencing datasets from DartCap that we received in 2017 actually comprised nearly 2400 loci, but we discarded hundreds either because they were uselessly uninformative (one very-dominant allele) or on QC grounds, e.g. possibility of paralogs— "better safe than sorry". This left us with 1541 loci that work as they should— close to our original plans. While people usually think of "SNP loci" as being purely biallelic, our loci are actually sequences of ~75-base-pairs, and many of the loci have several mutation sites within that sequence. To keep computations feasible, though, we chose only two "allowed alleles" (sequence variants, arbitrarily labelled A and B) at each locus.

Null alleles[2] are very common in the SBT genome, and this would normally reduce the power to detect kin-pairs since homozygotes (XX) are conflated with single-null (XO) genotypes. However, read-depths are so high with DartCap (typically several hundred reads per locus) that XO can usually be distinguished from XX based on the total reads of allele X for that sample-and-locus. This means that the nulls are serving as a 3rd allele, actually *increasing* the power to detect kin. Given the large number of samples needed for SBT CKMR, we would need considerably more loci (and expense) without this refinement. The idea is explained at greater length in Bravington et al. (2015).

Of the 1541 loci, we are using 1484 genotyped as just described to 6-way level (AB/AA/AO/BB/BO/OO). For the other 57 loci, the XX/XO separation did not work reliably, so we genotyped only to 4-way level, i.e. AB/AAO/BBO/OO where XXO means "either XX or XO". Examination of replicates indicates that error rates among the 4-way categories are very low— well under 1%. For simplicity, the POP-finding step uses only the 4-way genotypes. The HSP-finding step requires the extra information in the 6-way genotypes, where the locus-specific error rates for XO/XX are substantially higher (10% is common), and due allowance needs to be made for the possibility of such errors.

[2]Some of the nulls arise our choosing only two variants at a locus, even if there are sequenceable 3rd, 4th, ... variants. However, most nulls in SBT are "genuine nulls"— particular sequence variant(s) that do not get sequenced at all. These are repeatable and heritable, and are presumably due to mutations in the restriction-sites targeted by ddRAD. Such nulls are not merely failures-to-observe a "normal" allele resulting from low read-depths, which can happen with some GBS methods.