Update on the close-kin genetics project for estimating the absolute spawning stock size of SBT

Mark Bravington
Peter Grewe
Campbell Davies

Prepared for the CCSBT Extended Scientific Committee for the 15th Meeting of the Scientific Committee 4-10 September 2010
Taipei, Taiwan

Abstract

This paper describes progress with the close-kin estimate of spawning biomass. There are now genotypes for about 5000 fish at up to 22 loci spanning 3 years of sampling in both Indonesia and the GAB, and we expect to have 7000 fish done by the end of 2010. The paper gives outcomes for basic feasibility checks: reliability of identifying parent/offspring pairs, and checking for excessive numbers of sibs or half-sibs.

Contents

List of Tables

1 Overview

This paper is an update on progress in the SBT Close-Kin Abundance project. The project began formally in 2008, though data collection started two years earlier. The *modus operandi* is to identify Parent-Offspring Pairs (POPs) via "DNA fingerprinting" (multilocus genotyping), amongst comparisons between adults caught on the Indonesian spawning grounds and juveniles caught in the Great Australian Bight. For samples of given size, the *expected* number of POPs is inversely proportional to the total number of spawning-age adults in the population. This fact can be used to turn the *actual* number of POPs found into a formal estimate of spawner abundance, and thus SSB, using mark-recapture principles. The estimate requires very few assumptions, is fishery-independent, and is not vulnerable to the reporting-rate issues that can plague conventional tagging programs. Background is given in Bravington and Grewe (2007).

Although the basic principles of the project are straightforward, a number of technical issues have had to be addressed. Most notable is the need to collect large numbers (7000+) of tissue samples and genotype them at large numbers of loci, in order to reliably identify enough POPs to permit model checking and to give a precise abundance estimate. The project is the first to attempt genotyping on such a large scale (for non-human subjects). The logistics and quality-control (QC), which are essential to the project, have required a great deal of work.

The project seems to be working. We have now genotyped over 4000 fish with sufficient reliability to eliminate false-positive POPs, and we have indeed found a number of definite POPs. We have checked the incidence of full- and/or half-siblings in the juveniles, and there seem to be too few to have any impact on CV. For reasons explained below, it would be quite inappropriate to estimate SSB using these very preliminary results. The number of POPs so far is also too small to look for phenomena such as skip-spawning. Nevertheless, the results are entirely encouraging, and consistent with the project schedule. We are on track to finish genotyping 7000 fish with full QC protocols this year, and to produce as planned an estimate of SSB for CCSBT 2011.

Extrapolating from the samples analysed so far, the number of POPs eventually found will probably be lower than we assumed when designing the project. This is not surprising, because the number found is dictated by the very quantities that are uncertain and that the project is in fact designed to estimate, in particular SSB. However,

Table 1: Status of samples

Year	Collected	DNA extracted	Genotyped
	Adults (Indonesia, Sep-Apr)		
2005-2006	216	216	
2006-2007	1520	1069	644
2007-2008	1594	1200	1130
2008-2009	1700	1700	736
2009-2010	1840*		
2010-2011	[1500]		
	Juveniles (Port Lincoln, Jul-Sep)		
2006	4000	600	478
2007	4000	800	736
2008	4000	1288	1104
2009	4000	1248	
2010	3300+		
2011	[4000]		
Totals	25230	8121	4828

[]: planned
+: ongoing
*: 640 still in Indonesia awaiting permits to export research samples, for which the rules are currently being changed.

fewer POPs would have two undesirable consequences. First, the CV of the SSB estimate would be higher; second, there would be less ability to check the POPs for phenomena such as skip-spawning, which if detected would necessitate adjustments to the estimation model. The resources of the current project extend only to genotyping the originally-planned 7000 fish, but we do have a further 18000 archived tissue samples (Table 1). Increasing the sample size by genotyping some or all of these would proportionally increase the number of POPs, and the confidence in the final result.

Separately to the current project, we are also continuing the collection of tissue samples from adults in Indonesia and juveniles in Port Lincoln. These could be genotyped in future to develop a time-series of SSB estimates.

2 History since CCSBT 2009

By CCSBT 2009, we had selected and optimized an initial set of loci, and had begun checking for sibship amongst juveniles. This year, we started by genotyping a subset of 500 juveniles from a single year and cohort at 11 loci, to check more thoroughly for high incidence of full- and/or half-sib(ling)s. If the sib incidence was very high, then the number of parents being looked for would be much less than twice the number of juveniles, and the CV of the estimate would be worse than expected[1]. In extreme cases, this could render a close-kin project infeasible, so it is important to check for high sibship levels before embarking on the expense of full-scale genotyping.

Results on locus development and sibship checking were presented to the project's Steering Committee in May 2010; there was no evidence of substantial sibship amongst the 500 juveniles (see also section 3.3). The Steering Committee agreed that the project should go ahead with genotyping the remainder of the planned 7000 fish, aiming to get 5000 done by CCSBT. The set of loci was to be expanded well beyond 11, to avoid swamping genuine POPs with false-positives POPs (see next section and Appendix).

Between May and July 2010, about 4800 fish were genotyped at 22 loci, and the data were entered by late July. After an initial (but not comprehensive) clean-up, a usable version of the dataset was ready by 10 August. The results in this paper are obviously preliminary.

Sample collection and preparation have continued throughout, as shown in Table 1.

3 Results of genotyping

The goal of the genotyping is to find all the POPs, and to do so without ambiguity. In particular, there should be a negligible proportion of false-positive POPs (unrelated pairs that happen to look like POPs), and enough

[1] As explained in Bravington and Grewe (2007), sibs and half-sibs do *not* cause bias in the abundance estimate, but can affect CV.

headroom to distinguish between false-negative POPs (true POPs which appear not to be, due to some error in the genotyping) and almost-false-positives (unrelated pairs that look like false-negative POPs). The solution to both issues is to use plenty of loci. More loci do require more resources, but the marginal cost of scoring a few extra loci is small compared to the difficulties and uncertainties that arise when a substantial proportion of POPs are likely to be false. The Appendix gives further background on false-positives, false-negatives, and criteria for assessing POP status.

Since 2009, we have greatly increased the number of loci used, because it became clear that we would otherwise encounter numerous false-positive POPs. We are currently using 22 loci, although one has proved hard to score reliably on a substantial fraction of the 5000 fish, and has been excluded from the routine checks. Of the remainder, seven show statistically significant evidence of null alleles (Appendix), but at low levels. Fortunately, it is easy to relax the exclusion criterion used for assessing POP status to allow for nulls, and there is little increase in the false-positive rate provided the null-allele rate is small.

Most samples have been scored successfully at most loci. About 600 fish (grouped into 6 Plates of 92 fish) are currently missing an entire Panel (a group of 3-6 loci that are all processed simultaneously). Those plate-panel failures are only temporary; they can be resolved cheaply by re-processing. Aside from those, only about 5% of the samples (228 of 4800) appear to be unusable, i.e. failing to score at large numbers of loci. These encouraging figures reflect the generally very high quality of the tissue and its state of preservation.

The volume of samples and loci is immense: 5000 tissue samples were cut down to size, and then had DNA extracted; DNA was organized into 54 plates each of 96 fish; each plate was processed 5 times, with a different subset of the 22 loci being scored each time; the final dataset contains 200,000 alleles. To avoid mixups, stringent QC is needed at all steps along the way. We are still finalizing the QC process, but by the end of 2010, it will be possible to trace each fish through every step, to confirm that the fish still has the genome it started with.

3.1 POPs

To avoid problems with false-positives, we are currently restricting attention to fish with at least 16 scored loci, which equates to about 2000 adults and 2000 juveniles. Because different fish have different loci missing, the number of loci compared in each pair varies between 10 and 21, so individual comparisons vary considerably in their false-positive probability. However, with this subset of fish, the total expected number of false-positives is just 0.03.

We found 7 POPs that matched at every locus compared. They look genuine; all were based on comparisons of 18 or more loci, and also matched at the extra 22nd locus, which was not used in screening for POPs. There were 2 pairs that failed to match at just one locus. On investigation, both seem to be almost-false-positives (unrelated), rather than false-negatives (true POPs with a mis-scored locus). For one thing, the numbers of loci compared were just 14 and 11, unlike the 18+ comparisons involved with the definite POPs, so the chance of an almost-false-positive is relatively high; also, we have re-checked their mismatching loci, and there is no ambiguity about the genetic signatures.

As a consistency check, to see whether the process would generate false POPs, we also compared the juvenile sample just with itself, and the adult sample just with itself. In each case, the number of comparisons involved is about the same as for the real juvenile-adult comparison. However, since there cannot be any genuine juvenile-juvenile or adult-adult POPs[2], any POPs found would be have to be false-positives. Reassuringly, there were none.

There are some subtleties around the criterion for deciding which fish (or pairs of fish) to use in comparisons. The 16+-loci criterion used so far is temporary, and we will refine it intersessionally. Table 2 shows the effect of increasing the stringency; moving down the rows, a clear gap opens up between true POPs in the leftmost column and the closest almost-false-positives to the right, but at the expense of sample size in the Total column. In terms of an SSB estimate, there would be a bias-variance trade-off: stricter criteria mean fewer false positives and thus less bias, but also lower sample sizes and thus increased variance. Once the criterion is stringent enough to keep the expected number of false-positives well below one, then there is no reason to make the criterion any stricter. For criteria based on minimum-number-of-loci, this occurs at the 16+ level, where the expected number of false-positives is 0.03. In the table, the 16+ row is also where a clear dip between complete matches and single mismatches first appears.

The 15+ row is interesting. There is no dip in the frequency of almost-POPs, suggesting that false-positives POPs are likely, and this is confirmed by probability calculations. There are an extra 7 apparent POPs compared to the 16+ row, but 6 of them compare only either 9 or 10 loci, and may be false-positives. The 7th extra potential

[2]After excluding comparisons of a fish to itself, obviously. Note that it *is* theoretically possible to have an adult-adult POP, but there would have to be an age gap of at least 10 years. The number of such comparisons is very small.

POP is a 15-locus comparison, and its status is unclear. Many of the exactly-15-loci fish will gain extra loci once the failed plate-panels have been re-run, so these particular mysteries will be resolved. On the whole, it seems that the current set of loci are almost sufficient to deal with the false-positive-and-negative issue but, as the Appendix explains, it will be more secure once an extra couple of loci have been added, particularly if larger samples of fish are compared in future.

Table 2: Number of comparisons.

		Number of mismatching loci						
		0	1	2	3	4	...	Total
Min .#loci for inclusion	15	14	35	310	1527	6157	...	4807224
	16	7	2	15	110	896	...	4096196
	17	7	0	6	62	462	...	3840489
	18	7	0	3	28	247	...	3486439
	19	7	0	1	12	106	...	2842784
	20	3	0	0	4	47	...	1964118

Rows are strictness of criterion; columns are number of mismatching loci. The zero-column shows POPs, be they genuine or false-positive; the one-column shows false-negatives or almost-false-positives involving a single locus; the two-column shows pairs that mismatch at two loci, etc. The Total column shows the "sample size".

3.2 Description of the POPs found

The parents have not yet had their otoliths read. That aside, here is a brief description:

- All 7 parents are distinct (i.e. no more than one matching offspring per parent).
- There are 2 female and 5 male parents.
- Most of the 7 parents were slightly above the median size of adults captured in the same year. Sizes ranged between 161cm/93kg and 177cm/115kg. The parents would of course have been smaller when they spawned their offspring.
- 3 parents were caught in 2006/7, 1 in 2007/8, and 3 in 2008/9.

There is less to say about the offspring. All were 3-year-olds, but then we have deliberately concentrated on genotyping 3-year-olds so far. One was caught in 2006, and six in 2008 which, as per Table 1, is also when the most juveniles were sampled[3].

Once sufficient POPs have been found, they can be used to detect and estimate various unexpected aspects of breeding biology, such as skip-spawning, relationship between size and residency on the breeding grounds, and temporal stock structure (e.g. if all parents of GAB juveniles were caught in a limited part of the spawning season). However, with only 7 POPs to date, it is impossible to say anything definite about these such questions yet.

3.3 Checking sibship

The genotype data can also be used to estimate the incidence of sibs and half-sibs within each cohort and capture-year of juveniles. As noted last year, the genotype data is much less informative about sibship, and particularly half-sibship, than about POPs. Studying sibship in large samples is is a hard problem statistically, much harder than the main part of this project. There are no existing statistical algorithms, so we have developed our own. The (extensive) technical details are omitted from this report, but have been reviewed by the Steering Committee.

Since May 2010, we have adjusted the sibship algorithm to cope with null alleles, which can otherwise give spurious evidence of sibship. We have also applied it to the entire juvenile sample (one year at a time), not just to the subset mentioned in Section 2. The results give no indication that sibship is common. The point estimate is zero in two years, and involves just a handful of fish in the third. Because the number of loci used is limited, the results cannot exclude the possibility of small numbers of sibs and halfsibs, but small numbers would in any case pose no problem for the project. Simulations suggest that our algorithm is effective at detecting high incidences of sibship, given the number of loci and the sample sizes used, so the absence of evidence can be taken as evidence

[3]This oversampling was deliberate, to increase the proportion of the adult sample that would have been mature when the juvenile samples were spawned. Appropriate allowances will be made in the full SSB estimation model.

of absence. It is notable that none of the POP parents match to multiple offspring, so sibs/half-sibs cannot be overwhelmingly common.

We will continue to analyse sibship as more samples are genotyped.

4 Discussion

The results so far are very encouraging, given the innovative nature of the project. However, there are a number of reasons why these interim results **should not be used in a formal abundance estimate**. The following list is not exhaustive:

- QC is still incomplete. In particular, there is a real possibility that some as-yet-unchecked plate-panels may have been inadvertently swapped, as actually happened for a couple of the plate-panels that we have been able to check (and correct). Every undetected plate-panel swap would affect about 200 fish, completely disguising any POPs in about 10% of comparisons so far, so it is crucial to finish those checks.
- It is not yet clear what the appropriate number of comparisons is for estimating SSB. For example, there has been no exclusion of those "adults" that would actually have been immature at the juvenile birthdate; a just-mature adult in 2008 could not be the parent of a 3-year-old juvenile that was caught in 2006 and therefore spawned in 2003.
- 7 POPs is far too few to decide on appropriate model structures, e.g. whether or not skip-spawning is commonplace.
- With this small number of POPs, the sampling variability is very high. The 90% confidence interval for the "number of matches that *should* be present" is $[4.0, 13.2]$ and this range would be proportionally reflected in the interval for any SSB estimate.

Notwithstanding the above, it is worth noting that 7 POPs from about 4,000,000 comparisons is within the range to be expected, based on the range of values for 2004-2006 SSB considered in the OM scenarios, and using the naive assumptions of the simplest possible mark-recapture model.

5 Prospects for 2011

In this final year of the current project, we plan to:

- finalize QC procedures, including rescoring missing panels and reorganizing some of the existing panels;
- add a small number of additional loci, to widen the gap between almost-false-positives and false-negatives;
- genotype another 2000 fish (50/50 adults and juveniles);
- identify POPs
- read otoliths of the parents;
- check for patterns in the POPs and develop the full estimation model accordingly;
- estimate SSB and associated CV.

Note that the SSB estimate will be retrospective to juvenile birthdates, i.e. roughly 2004-2006. The precise definition will depend on details of the estimation model, e.g. whether it is time-averaged or disaggregated by year. Those details will not be decided until we have checked for patterns in the POPs.

The CV of our estimate in 2011 will depend primarily, though not exclusively, on the *number* of POPs that we find. All else being equal (sample sizes etc.), a smaller SSB would mean more POPs and a smaller CV. If the SSB was much larger than we assumed when designing the project— and the whole point of the project is that the true SSB is *not* known precisely— then the CV would also be larger[4].

The CV will also depend on any adjustments that need to be made to the estimation method following model checking. In this instance, "model checking" means looking for patterns among the POPs: e.g., if successful parents tend be unusually large, or if there is evidence of skip-spawning. Such phenomena would not present any fundamental

[4] Happily, the lower confidence limit would still be larger, despite the increase in CV.

problem for the project, but would necessitate adjustments in estimation which would affect CV; see Bravington and Grewe (2007). To do the model checking, there must of course be a reasonable number of POPs in the first place.

So far, we have used about 4000 fish in pairwise comparisons. Once we have 7000 fish genotyped, the number of comparisons will actually triple, thanks to the quadratic relationship with sample size. Extrapolating from our 7 current POPs, we might therefore expect around 20 POPs next year, although the final figure might be considerably higher or lower. Using the simplest estimation method, 20 POPs would yield a 22% CV, which sounds respectable. However, the estimation method that ultimately gets used is sure to be more complicated, since it will have to take into account the multi-year nature of the study, etc. The extra parameters required will increase the CV. In addition, the amount of model-checking that is feasible with 20 POPs would be limited.

A particularly important check is on the strength of any relationship between fish size and residency (length of time spent) on the spawning grounds. Residency affects not just catchability and therefore a fish's chance of appearing in the adult sample, but also the number of juveniles it is likely to have contributed three years previously (since most big fish now were also big three years ago). In mark-recapture terms, this induces "heterogeneity of capture probability". If the size-duration relationship is strong, it can bias the abundance estimate, and also the interpretation of SSB. As described in Bravington and Grewe (2007), it is possible in principle to estimate and allow for the size-duration relationship, by comparing the size distribution of parents with that of the general adult sample. A reasonable number of POPs will be needed to do this with confidence. Another way to infer this relationship, and an independent check on this aspect of the SSB estimation model, would be via archival tag recaptures from mature fish. However, this would take some years. In the meantime, genotyping some of the 18000 archived samples would be the fastest and cheapest way to improve model-checking and reduce CVs.

6 Acknowledgements

This project relies on our samplers. In Indonesia, particular thanks are due to Craig Proctor (CSIRO), Ms Retno Andamari (Research Institute for Mariculture, Gondol, Bali), Mr Kiroan Siregar, and Mr Rusjas Mashar (sampling enumerators) for their continued diligent efforts in ensuring the successful collection of tissue samples. In Australia, Tony Jones of Protec Marine has done an outstanding job in collecting over 19000 tissue samples.

We appreciate the assistance of the Australian SBT industry (Port Lincoln), the Indonesian tuna fishing industry (Benoa), and the Research Centre for Capture Fisheries (Jakarta).

Funding for this work is provided by CSIRO Wealth from Oceans flagship and by the Fisheries Research and Development Corporation.

References

Bravington, M. and Grewe, P.: 2007, A method for estimating the absolute spawning stock size of sbt, using close-kin genetics, *Scientific Committee Report CCSBT-SC/0709/18*, Commission for the Conservation of Southern Bluefin Tuna.

7 Appendix: how to identify POPs from DNA

Every fish has two *alleles* (i.e. particular sequence of DNA) at each *locus* (i.e. specific place in the genome). The loci we use are *microsatellite repeats*, where many different types of allele at each locus are found across the population. Usually, a fish has two different alleles at a locus (i.e. it is *heterozygous*), but by chance the two alleles may be the same (*homozygous*). The different types of allele are distinguished by their length, and *scoring the locus* means "measuring the lengths of the two alleles". *Genotyping a fish* means "scoring it at all loci possible"; sometimes a tissue sample simply will not yield a result at a given locus, but this does not matter as long as enough other loci are successfully scored.

An offspring inherits one allele at each locus from each parent, so a POP must *match*, i.e. have at least one allele in common, at every locus. The following *exclusion criterion* can therefore be used to determine POP status: two fish are deemed a POP if they match at every locus, but the presence of even one mismatching locus excludes the possibility of being a POP. Matches at all loci can also happen by chance even for unrelated fish; this would be a *false-positive* POP. However, the per-fish-pair false-positive probability becomes vanishingly small if enough loci are compared: about 2×10^{-10} for a 20-locus comparison in this project. Since the overall number of fish-pairs will only

be about 10^7, complete-match false-positives will not be a concern provided we restrict attention to comparisons with enough loci.

The picture is made more complicated by the possibility of *false-negatives*, i.e. true POPs that somehow appear to have one locus without an allele in common, and that therefore fall foul of the exclusion criterion. False-negatives could arise by either of two mechanisms. The first, *mutation*, can be ignored because of its rarity; estimates are typically $\sim 10^{-3} - 10^{-5}$ per generation. The second is *scoring error*, where the alleles recorded are not the same as the alleles actually present. There are many reasons for scoring error, but its incidence can be minimized by careful choice of loci, optimization of the associated scoring process, and having high-quality samples to begin with. Unlike false-positive rates, scoring error rates cannot be predicted; in this study, they will only be directly estimable after a large number of POPs have been identified. While typical estimates of scoring error in other studies are around 1% (i.e. at least 10 times as common as mutation), we can expect to do rather better, because of high quality samples and careful design. To be cautious, though, if the scoring error rate was as high as 1% per locus, then with over 20 loci scored, a significant fraction of true POPs would mismatch at one locus and be incorrectly excluded. (The proportion mismatching at two or more loci should be negligible.) We do therefore have to somehow relax the exclusion criterion to allow for the possibility of some false-negative POPs. These false-negatives could be confused with *almost-false-positives*, i.e. unrelated fish that fail to match only at one locus.

There are not many pairs that mismatch at just one locus, so they can be re-examined; this can sometimes identify a scoring error, and thus distinguish a false-negative from an almost-false-positive. However, scoring errors cannot always be detected on re-examination. The most secure solution is to further increase the number of loci, so that the expected number of almost-false-positives (one mismatching locus), as well as complete false-positives (zero mismatching loci), becomes negligible. The expected number of false-negatives increases linearly in the number of loci, whereas the expected number of almost-false-positives decreases exponentially, so with enough loci it becomes almost certain that any out-by-one comparison is really a false-negative. The almost-false-positive rate in this project is about 40×higher than the complete-false-positive rate, at around 10^{-8}. Hence, the current set of loci is just about adequate for our intended 10^7 comparisons, provided almost all fish are scored at almost all loci. The planned addition of another two loci should fully resolve any ambiguities about pairs that mismatch at just one locus, and should future-proof the comparisons as and when sample sizes increase.

The above omits several technical details, two of which are worth noting here:

- Certain *null* alleles at some loci may not be scorable, even though heritable in the usual way. Fish with a null allele will look like homozygotes with two copies of whichever other allele they have. This phenomenon can be detected statistically at a population-wide level, and allowed for by a slight relaxation of the exclusion criterion. A similar remark applies to a phenomenon called *long-allele dropout*. There are several different ways to handle these phenomena in the context of POP-hunting. Statistically, the most powerful approach looks to be a relaxation of the per-locus exclusion rule to accommodate those two phenomena only, plus a relaxation of the overall exclusion rule to allow (or at least re-check) single-locus mismatches, the latter being a guard against all types of scoring error.

- For, say, aunt-nephew comparisons, the complete-match probability also turns out to be very small, although much higher than for unrelated pairs. However, there are vastly more unrelated pairs than non-parent-offspring close-kin, by a factor of at least 10^5. Provided enough loci are used to eliminate false-positives from unrelated pairs, there will not be a problem with false-positives from non-parent-offspring close-kin.