CCSBT-ESC/1107/19

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 16th Meeting of the Scientific Committee 19-28 July 2011
Bali, Indonesia

CLOSE-KIN UPDATE FOR CCSBT SC 2011

MARK BRAVINGTON, PETE GREWE, CAMPBELL DAVIES: CSIRO HOBART, JULY 2011

Introduction

This paper gives an update on progress and plans in 2011 for 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 CCSBT-SC/0709/18.

Last year's update, CCSBT-ESC/1009/Info 2, described the genotyping (AKA scoring) and POP matching of about 4000 fish[1]. Seven definite POPs were found. Finding those POPs showed that the project was succeeding at a technical level. However, for the number of fish examined in 2010, the number of POPs found was substantially less than expected. In itself, this is not particularly surprising; as with many sample design problems, the actual sample size required to achieve a given precision (which in this case is set by the number of POPs) depends on the very thing that one is trying to estimate (in this case adult abundance), which is of course unknown. Nevertheless, it was clear in 2010 that, if the final sample size stayed at the 7000 originally planned, then there would likely not be enough POPs to give a precise estimate by the end of the project. The real issue is not so much the "CV" per se, but rather the need to get enough POPs to see any important patterns in them and then to formulate an appropriate and unbiassed statistical model accordingly. Even though the basic principles of POP-based abundance estimation are very simple, the application to SBT does require some care, because of the multi-year nature of the study and the interaction between the biology and sampling (fishing). For example, SBT may turn out to exhibit skip-spawning, and/or there may be links between adult size and effective fecundity. Both phenomena can be detected from POPs, and then allowed for in the statistical model, but only if there is a reasonable number of POPs in the first place.

In late 2010, the funding agencies CSIRO and FRDC therefore agreed to double the sample size (i.e. number of fish genotyped) to around 14-15,000. Extrapolating from the 2010 results, this should increase the number of POPs to[2] around 70-80, close to the original intention. The modification was possible because the project had deliberately collected a buffer of extra samples every year since 2006; the marginal cost of collection (as opposed to genotyping, which might never have been required) was low, and the extra samples provided a buffer in case the sample size ever did need to be increased. Even though the samples were already available, it has required a huge effort to process, genotype, and database 7000 samples within 8-9 months. Other tasks completed along the way (see below) have been to extend the set of loci (to give better surety about parent-offspring ID, and to cope with the increased potential for false-positives caused by an increased sample size), to fill in the gaps in the 2010 results, and to set up robust and traceable quality-control checks.

Timeline and further work. We are currently finishing a more formal range of quality-control checks to handle the greatly increased volume of genotyping data, which is now coming in at the rate of about 2000 fish per month and has undergone a number of changes of format since the start of the project. The files containing genotype scores are being linked to the existing CSIRO database, so that (adult) fish can

[1] Partial genotypes were available for another 1000 fish last year, but those data were too sparse for POP-hunting.

[2] This extrapolation is also uncertain, since it is based on a count of just 7, but is based directly on real data.

be cross-referenced for age, date of capture, etc. The QC code does need to be in place before another round of POP-checking can be done, but POP-checking itself is quite quick and straightforward. Once POPs have been found, most of the adult POP members will need to have their otoliths specially read, although some will already have been read in the existing Indonesian ageing project. These processes can happen in parallel with the rest of the genotyping, which we expect to have completed by mid-September (although delays in obtaining samples from Indonesia may compromise this slightly; see below). The statistical model will also be ready by the end of August, at least in preliminary form; its construction has deliberately been on hold until enough new data became available to select a sensible model. The draft final report is due on 30th December 2011.

No formal arrangements have yet been established for continued sampling from Indonesia or Port Lincoln in 2011/12, but the low cost of collecting the samples (even if they end up never being genotyped) presents a good case for continuation. Close-kin studies have the remarkable property of a *quadratic* gain in efficiency with sample size; extra samples now will both enhance the effectiveness of the existing data, and open the door to the development of a fully time-dependent (time-series) estimator of abundance.

Progress with genotyping

Table 1 shows the status of DNA extraction and genotyping for samples from various years of the study. We have now genotyped nearly 9000 fish. The whole procedure is now highly streamlined, with a mixture of in-house and outsourced steps to get the best in both quality control and cost-effectiveness, and the current throughput is over 2000 fish per month.

There is no formal design to our choice of samples from different years, and nor would it be possible to come up with such a design before the results are in. However, we aim for (i) a roughly even split between juveniles and adults, since that maximizes the expected number of POPs for a given expenditure (and in practice means using as many adults as possible, since they are in short supply), (ii) a fairly even spread across years, to give the best chance of allowing for any time-related effects that emerge, and (iii) selection of juveniles of fixed, known age (based on length). Recent changes to Indonesian administrative requirements for export of biological samples have led to unexpected delays in obtaining the necessary approval to ship the most recent samples, and in the worst case we may not have enough time to genotype them within the current study. However, even if that does happen, the shortfall can be made up from the stocks of juvenile fish.

The suite of loci has been expanded and reorganized since 2010. A total of 25 hypervariable microsatellite loci are now used for scoring and one for cross-checking, five more than in 2010. The key is to use enough loci to ensure that the expected number of false positives is much lower than the expected number of true positives, now expected to be 70-80. There will be about $7000^2 \approx 5 \times 10^7$ adult-juvenile comparisons in all, so the average probability of a false-positive needs to be kept to less than, say, 10^{-8} to expect less than one false-positive overall (number of comparisons times the probability of each one being false-positive). Not all loci are scored successfully for all fish, and a typical comparison will involve about 18 loci. The false-positive probability based on comparing 18 "typical" loci is about 7×10^{-10}, so there should be plenty of buffer against false-positives. The overall sample size may need to be reduced somewhat to exclude fish with few successfully-scored loci, since such fish will otherwise greatly increase the false-positive rate, but overall there should be plenty of loci even if the close-kin study continues and the sample size grows over time.

To check the consistency of our genotyping, this year a number of plate/panels have been re-scored (genotyped) by different readers. While there are very few instances of dramatically different scorings (say, one reader scoring as AB and another as AC or even CD), there are a modest number where a locus has been scored AB once and AA once. So far, it seems that that a genuine second allele has been overlooked in these cases, rather than a spurious allele being invented. Most of the fish have only been scored once, for obvious reasons of cost, and it is therefore likely that there are a few mis-scored "homozygotes" among them. The overall effect cannot be large, since scored homozygotes, whether real or not, are uncommon in all our loci (which are deliberately chosen for hypervariability, and thus low homozygosity). Nevertheless, it is possible that the low but non-zero estimates of heritable-null allele frequency at some loci, mentioned last year, are at least partly an artefact of genotyping error.

Table 1. Status of sample collection and processing, July 2011

Year/ Place	Collected / Archived	DNA extracted	Genotypes complete	Genotypes planned[†]
Adults (Indonesia)				
2005-6	216	216		210
2006-7	1520	1520	1520	
2007-8	1594	1594	1564	30
2008-9	1637	1632	1380	252
2009-10	1200+680*	1172	1104	748
2010-11	~1000*			~1000
Total A	**~7850**	**6134**	**5568**	**~2250**
Juveniles (GAB)				
2006	4010	1440	460	920
2007	4065	1472	736	644
2008	4027	1452	1104	276
2009	4103	1440	1012	368
2010	4071	1440		1380
2011	[4000]			
Total J	**24300**	**7244**	**3312**	**~3550**
TOTAL	**32850**	**13378**	**8880**	**~5800**

* : in storage awaiting export from Indonesia

[†]: approximate, depending on best way to organize genotyping

To illustrate how we plan to handle POP-finding and the issues around false positives, false negatives, and scoring error, we include an exact copy of last year's Table 2; note that it has not yet been updated to include the new data. With more loci scored this year, the gap between the "lucky lookalikes" (the right-hand columns— pairs of fish which by chance share alleles at a lot of loci) and the true POPs (the left-hand column) will be bigger than in the Table. Even if the occasional scoring error does results in a small number of pairs with an apparent mismatch at one locus (i.e. true POPs which should be in the left-hand column, but has moved into the next one), they will still be clearly separated from the lucky lookalikes, and will not be automatically rejected— i.e. they will not become false negatives. Any pairs will small numbers of mismatching loci will be re-scored, and their POP status will be assessed taking into account the nature of the apparent mismatch (as per previous paragraph).

Also, we have now filled in most of the big gaps from last year. These arose when an entire plate/panel failed to work, thus removing about 5 loci from 100 fish and leading to a lot of pairwise comparisons involving rather few loci. This means we can now afford to be reasonably stringent about the "entry requirement" for a comparison (i.e. how many loci must be scored in both fish of a pair, for that comparison to be considered) without sacrificing too many potential comparisons. In terms of Table 2, that means we should have a good-sized "sweet spot" between the bottom row (where the entry requirement was too strict, and the number of comparisons was thus much reduced) and the top rows (where the entry requirement was too lax, and the lucky lookalikes overlapped with the true POPs).

We have also tuned our quality-control procedures this year. With so many fish involved, and each plate of ~96 fish needing to be run through the equipment on several separate occasions (i.e. in different "panels"), it is essential to have some way to check that the fish have not become muddled up. Each plate of fish is uniquely coded by using water "blanks" in specific positions. Each plate also uses positive controls (i.e. two known "standard" fish) in specific positions, to give a controlled product and an additional key to identify a plate across all panel runs. All runs are multiplexed, with a standard tube used for each panel and all plates. This and other protocols have allowed us to detect and fix several problems that could otherwise compromise identification of POPs.

Table 2. Number of pairwise comparisons, by number of mismatching loci involved (COPY OF 2010 TABLE, WITHOUT NEW DATA; JUST FOR ILLUSTRATION). Rows are strictness of entry requirement for a pair of fish to be compared, in terms of the minimum number of loci at which *both* fish are successfully scored. Columns are number of mismatching loci with. 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", i.e. total number of pairs that meet the entry requirement.

		Number of mismatching loci						
		0	1	2	3	4	...	Total
	15	14	35	310	1527	6157	...	4807224
Min .#loci	16	7	2	15	110	896	...	4096196
for inclusion	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

Acknowledgements

This project relies on our samplers and technical support crew. In Indonesia, thanks are due to Craig Proctor (CSIRO), to the scientist team at Research Institute for Tuna Fisheries (Benoa, Bali), and in particular to 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 and Adam Kemp of Protec Marine have done an outstanding job in collecting over 20000 tissue samples from Sam's Tuna and Tony's Tuna. At CSIRO Hobart, Peta Hill, Rasanthi Gunasekera, Matt Lansdell, Scott Cooper, Danielle Lalonde, Bruce Barker, and Mark Green have made heroic efforts this year to process, genotype, and database what has become an enormous quantity of samples.

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

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