Report of the Close-Kin Project: estimating the absolute spawning stock size of SBT using genetics

CCSBT-ESC/1208/19

Mark Bravington Pete Grewe Campbell Davies

CSIRO Marine Lab, Hobart: July 2012

**In confidence to CCSBT SC;
not for circulation**

Contents

Abstract

It is in principle possible to estimate the absolute abundance of adult SBT without using catch or CPUE data, via a variant of mark-recapture applied to parents and offspring identified by genotyping large numbers of adults and juveniles. The method was first described in CCSBT-SC/0709/18, and since 2006 we have been running a large project to implement it. The project is now coming to a successful end, and this paper describes the [main] outcomes. We genotyped over 13,000 SBT caught between 2006 and 2010 in the GAB (juveniles) and off Indonesia (mature adults), and found 45 Parent-Offspring Pairs (POPs). Combining data from the POPs (the number found, plus their age, size, sex, and date of capture) with fecundity-at-size studies and Indonesian length, sex, and age-frequency data, we constructed a self-contained assessment of absolute adult abundance that does not require any catch or CPUE data. As well as abundance, we were able to estimate adult survival, selectivity-size relationship, and effective female reproductive contribution as a function of length. This paper explains the method, and presents an example of results for a steady-state scenario. These results, plus those from a limited number of other scenarios explored to date, indicate that adult abundance is considerably higher than current OM estimates. A small amount of work remains to finalise our self-contained assessment and more fully explore the model uncertainties, and we expect to complete it in the coming months as part of final project reporting. Options for the integration of the new data into the OM are considered in CCSBT ESC 1208/21.

Acknowledgements

This project has run for over 5 years with contributions from many people, too many to thank individually. Particularly heroic efforts on genotyping and logistics have come from Peta Hill, Rasanthi Gunasekara, and Matt Lansdell of CSIRO, with Jess Farley of CSIRO providing invaluable help with otolith and fecundity data. Special thanks for their diligent efforts in ensuring the successful collection of Indonesian tissue samples are due to Craig Proctor (CSIRO); Mr Kiroan Siregar, Mr Rusjas Mashar, and other scientists at the Tuna Fisheries Research Institute of Benoa; and Ms Retno Andamari (Research Institute for Mariculture, Gondol, Bali).

We greatly appreciate the continued assistance of the the Australian SBT industry (Port Lincoln), the Indonesian tuna fishing industry (Benoa), and the Research Centre for Fisheries Management and Conservation (Jakarta). Funding has been provided by CSIRO Wealth from Oceans flagship and the Fisheries Research and Development Corporation of Australia, with in-kind support from the Ministry of Marine Affairs and Fisheries, Indonesia.

1 Introduction

The SBT close-kin abundance project rests on two simple ideas:

- modern genetics allows us to tell whether any two fish constitute a Parent-Offspring Pair, via "paternity analysis";
- all juveniles have two parents.

Consequently, if you compare any juvenile with a randomly-chosen adult, there is a probability $2/N$ that you will discover that the adult is one of the juvenile's parents, where N is the adult population size. Given large samples of m_J juveniles and m_A adults, and repeating the comparison for each of the $m_J \times m_A$ pairs, some number P of Parent-Offspring Pairs (POPs) will be observed. The expected value of P is $2m_J m_A/N$, so using the observed value the adult abundance can be estimated from $\hat{N} = 2m_J m_A/P$. The devil lies in two details: doing the genetics well enough to correctly find the number of POPs, and adjusting for sampling biasses in the "randomly-chosen adults" aspect (e.g. selectivity biasses towards larger fish).

The project began collecting adult samples in Indonesia during the 2005/2006 spawning season (November-April), and in Port Lincoln during the 2006 season. Subsequent progress has largely followed the approach described in CCSBT-SC/0709/18, and a progress report has been given to CCSBT each year. Preliminary results presented to CCSBT in 2010 and 2011, after most of the originally-planned sample size had been genotyped, showed that the number of POPs found would end up considerably lower than originally expected if the original sample size was maintained, which would make the precision of the final results very uncertain. It was therefore agreed to substantially increase the sample size. This was straightforward since we have available many more frozen tissue samples from juveniles 2006-2010 than we have budget to genotype, but of course required extra time, so the final results of the project are only available now. Oversight of the project has been provided throughout by a Steering Committee including expertise in genetics, mark-recapture, and SBT assessment.

The main departures from the original project plan are that we have genotyped more fish over a longer period, used more loci to ensure the genetics are unambiguous, and have developed a full stand-alone statistical mini-assessment of adults during the 2000s in order to deal with the complexities of growth, selectivity, and fecundity. The assessment uses the POP data, external studies on fecundity-size relationships, and Indonesian length/sex/age composition data, but no catch or CPUE data.

Table 1 shows the final breakdown of 13,023 genotyped samples by year and site. A few hundred more were genotyped, but excluded in the end for assorted quality-control reasons. Although the optimal scheme for a given budget would have been to genotype equal numbers of juveniles and adults (since this is likely to yield the greatest number of POPs for a fixed amount of genotyping effort), regulatory changes and delays with Indonesian export permits meant that we had to shift the balance somewhat towards juveniles. Almost all the Port Lincoln juveniles were age 3 in the year of sampling (based on clear separation of modes in the length frequency), except for a few in 2006 that were age 4. After 2006, the Indonesian samples were taken from every available fish (almost all >150cm length) alongside the existing length/weight measurement and otolith-collection schemes[1]. Sample collection is continuing in both Indonesia and Port Lincoln, but there are no immediate plans or funding to genotype more samples; they are simply being frozen for possible future use.

[1] In two years, some Indonesian vessels fished further south than usual, off the main spawning ground, and were catching subadult fish. Fish from those vessels and years are excluded both from the genetic study and from the length-frequency data.

Table 1: Final tally of fish genotyped successfully. For Indonesia, "year 2006" means "spawning season from November 2005 to April 2006", consistent with the definition of "SBT birthdays".

	2006	2007	2008	2009	2010	Total
Indonesia	214	1457	1526	1394	1164	5755
Port Lincoln	1523	1707	1448	1338	1432	7448
Total	1737	3164	2974	2732	2596	13203

In the rest of this report, section 2 describes the principles and the results of our quest for POPs; further details of the genetic procedures and QC aspects are given in Appendix 1 (section 6), and further statistical details in Appendices 2 and 3 (sections 7 and 8). Section 3 describes some qualitative findings about the POPs found (e.g. average size), section 4 explains the mini-assessment and shows results from one example of its application, and section 5 is a summary.

2 Genetic results: finding POPs

We use the genetic data to find POPs, by first genotyping all the fish and then comparing every juvenile to every adult, eliminating non-POPs via "Mendelian exclusion" as described next. A brief guide to terminology can be found at the start of the Appendix, which contains a more detailed description of the operational aspects of genotyping.

Every animal has two alleles at each locus, though the two may by chance be the same; one is inherited from each parent. Therefore, a POP must share at least one allele at every locus. If there are one or more loci at which the pair do not share an allele, then the pair is not a POP. Although two non-POP individuals could by chance share an allele at every locus compared, the probability is very low if the number of loci examined is large and the loci are individually highly variable, so that no one allele is particularly common. Therefore, the most basic and most rigid exclusion principle is: a pair is treated as a POP if and only if the two animals have at least one allele in common at all loci.

This project relies on the number of POPs actually identified being close to the true number of POPs in our samples. There are two possible issues. The first is false-positives: an unrelated pair might happen to share an allele at every locus just by chance, and thus look like a POP. This probability can be assessed in advance from the allele frequencies, and this step is essential in determining whether enough loci are being used. Not all loci are successfully scored for all fish, so some comparisons will involve a lot fewer than the theoretical maximum of 25 loci in our study, and those comparisons will have a substantial false-positive probability. By excluding those "weak" comparisons, we can control the overall false-positive rate so that the expected number of false positives is negligible compared to the number of true positives[2].

The second possible issue is false-negatives, whereby a POP *appears* not to share an allele at one or more loci. This could arise through mutation, but only very rarely; published estimates of mutation rate for the kind of loci that we used are of the order of 10^{-4} per generation, so with about 25 loci in our comparisons well under 1% of true POPs would be affected by any mutations. A more likely cause of false-negatives is scoring error, whereby the true alleles are incorrectly recorded. Scoring error rates are highly variable between studies (and to some extent between loci within a study), depending on the quality of the DNA itself (i.e. tissue

[2]When we originally planned the study, our intention was to use fewer loci in the first pass (about 15), but to have a suite of "backup" loci that could be used to double-check possible POPs. However, both on economic grounds and more importantly to minimize the chance of processing mixups whereby the wrong animal gets scored, we decided to use many more loci (about 25) and to deploy them all on the first pass.

preservation), how carefully the loci are chosen, how carefully protocols are followed, and how much checking is done. Careful checking can detect and eliminate large-scale scoring errors involving many fish at once (see 6.5). However, a different approach is required for small-scale errors at the level of single loci on single specimens.

Because there are so many different possible causes of scoring error, false-negative rates cannot be predicted in advance (unlike false-positive rates), and can only be inferred after the fact. This is usually done by re-scoring individuals to see how often the scores change. However, depending on the details there may be a possibility of making the same mistake twice, so re-scoring may underestimate the scoring error rate. With our POP-oriented study, we can use a more direct and robust approach; we are using so many loci that the chance of two non-POPs sharing an allele at all-but-one of (say) 25 loci is negligible, and consequently any pairs that *seem* to share alleles at 24 of 25 loci with a mismatch at the 25th are highly likely to be false-negatives arising from scoring error[3]. The proportion of such cases compared to unambiguous true POPs (where all loci share an allele) can be used to estimate the overall false-negative rate.

2.1 False positives?

Barring errors, a POP must have at least one allele in common at every locus, so if a pair is unrelated we will eventually be able to rule it out as a POP by finding a locus that does not share an allele, provided that we look at enough loci. We have scored 25 loci[4] overall, but not all loci are scored for every fish, so some pairwise comparisons involve many fewer loci. If too few loci are used in a comparison between unrelated fish, there is a substantial probability that all the loci will share an allele just by chance. We therefore need to do some filtering, to exclude comparisons that are too likely to give a false positive. Table 2 shows what happens if we *don't* do any filtering. True POPs— plus false POPs, which just happen by chance to share an allele at every locus compared— are in the leftmost column "F0", i.e. with zero loci compared that do not share an allele. False POPs are obvious in the top-left of the table, where very few loci are being compared.

Note that the Table includes a small proportion of (i) impossible and (ii) useless comparisons, where the adult was (i) caught in a year before the juvenile was born, or (ii) caught in the same year. Type (ii) comparisons are biologically possible, but it's not helpful to include same-year comparisons in abundance estimation, because in the year of its capture an adult will not achieve its normal annual reproductive output. All such comparisons have been removed in subsequent summaries and results.

[3]Note that scoring errors do not increase the false-positive probability— there is no reason to think that an error in scoring one fish will either increase or decrease the probability of it sharing an "allele" with another unrelated fish.

[4]Plus another two that showed occasional anomalies, and were therefore omitted from routine pairwise comparisons, but were used in checking ambiguous possible-POPS.

Table 2: All comparisons, broken down by #loci compared and #loci inconsistent with POPhood. Hash (#) means "number of", dot means zero, plusses mean too big to fit.

Table 1: tab:all-comp-by-fail Breakdown of all comparisons, by #loci and #excluding loci

.	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20	F21	F22	F23	F24	F25	TOTAL
C0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0
C1	9435	19641	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	29100
C2	6400	26740	22968	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	56100
C3	1785	7894	12047	6297	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	28000
C4	997	6612	14848	15184	5568	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	43200
C5	273	2832	10961	20350	18536	6635	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	59600
C6	83	927	4581	12245	18839	15132	5005	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	56800
C7	42	599	3287	9903	18591	20576	12813	3405	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	69200
C8	7	173	1309	5349	13521	21186	20628	11667	2745	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	76600
C9	1	54	554	2567	8278	17043	22652	18913	9242	1929	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	81200
C10	3	54	469	2454	8868	21055	34780	39403	29174	12807	2434	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	152000
C11	1	34	320	1933	7983	22325	45367	64193	64783	42695	16702	2815	.	.	.	.	.	.	.	.	.	.	.	.	.	.	269000
C12	.	16	208	1307	5488	16855	37136	61101	71266	60071	33801	11794	1808	.	.	.	.	.	.	.	.	.	.	.	.	.	301000
C13	.	7	104	860	4089	13833	35071	65182	90261	89534	64433	30674	9185	1190	.	.	.	.	.	.	.	.	.	.	.	.	404000
C14	1	4	74	643	3237	12411	35917	76584	+++++	+++++	+++++	92018	40818	11013	1342	.	.	.	.	.	.	.	.	.	.	.	695000
C15	.	3	42	383	1998	8509	27430	66904	+++++	+++++	+++++	+++++	99274	40608	10210	1144	.	.	.	.	.	.	.	.	.	.	927000
C16	2	1	18	131	966	4716	17097	47526	+++++	+++++	+++++	+++++	+++++	+++++	40747	9691	1058	.	.	.	.	.	.	.	.	.	1170000
C17	2	.	8	92	655	3674	14677	45482	+++++	+++++	+++++	+++++	+++++	+++++	+++++	48659	10815	1112	.	.	.	.	.	.	.	.	1940000
C18	5	.	6	65	483	2699	12037	40524	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	56338	11998	1162	.	.	.	.	.	.	.	3060000
C19	7	.	1	33	288	1728	7992	29511	87021	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	55653	11047	1030	.	.	.	.	.	.	4160000
C20	2	1	1	15	131	886	4630	18722	60834	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	54641	10390	946	.	.	.	.	.	5510000
C21	14	.	1	5	62	481	2589	11387	40151	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	52231	9298	788	.	.	.	7200000
C22	.	.	.	.	4	38	165	698	2737	8877	23841	53778	+++++	+++++	+++++	+++++	+++++	+++++	67828	27586	7668	1364	117	.	.	.	1170000
C23	4	.	.	.	2	20	143	754	3402	11715	34645	84564	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	51985	14118	2383	179	.	.	2970000
C24	2	.	.	.	4	22	90	558	2596	10110	31919	85623	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	68298	17376	2799	214	.	5100000
C25	6	.	1	.	1	5	22	199	910	3747	13071	38100	94463	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	+++++	42419	10339	1607	139	4120000

In order to filter out false POPs, we first compute in advance for each possible pair a False-Positive Probability (i.e. the probability that the two animals will share an allele at every locus compared, even if unrelated) based on which loci were scored successfully for *both* fish in the pair, and without looking at the actual genotypes that resulted. We then sort these FPP in ascending order, and find the cutoff such that the *total* FPP from all (sorted) pairs below the cutoff is below some pre-specified threshold T. Only those pairs whose FPP falls below the cutoff are subsequently checked for POPhood, the remainder being deemed too ambiguous. Note that not testing POPhood of an ambiguous pair does not cause any bias in the proportion of included comparisons that yield POPs, because the FPP check is done *before* testing for POPhood, and is unrelated to whether the pair really is a POP or not. The threshold T is by definition equal to the total expected number of false POPs, so we choose it to be a small fraction of the number of true POPs, of which we have a shrewd idea of by this stage. For this report, we have set the threshold at 0.35, below 1% of the number of POPs actually found. Because false POPs lead to a proportional negative bias in abundance estimates, the upshot is that we have kept such bias to under 1%.

The resulting set of filtered comparisons is shown in Table 3. At least 11 loci must be compared to get an FPP above the cutoff, and less than 100 11-locus pairs squeeze in; these occur where the 11 happened to be amongst the most powerful[5] of the 25 loci used for the table. On average, the loci used have about a 0.65 chance of *not* sharing an allele by chance, and the table shows very clearly how (near-)binomial probabilities work; from right to left, the numbers in the columns decline rapidly, except for the leftmost column where true POPs appear.

Importantly, in the bottom-left-hand-corner, the Table shows "clear blue water" between the best-matching unrelated pairs (i.e. with fewest loci that do not share an allele) and the true POPs. The separation is less obvious in the rows above say C16, but by looking at how fast the numbers in each row decline from right to left through the F4-F3-F2 columns, it is clear

[5]I.e. genetically more diverse, and being least likely to share an allele by chance

that very few unrelated pairs would have made it into the F0 column. And of course this is what the FPP calculations suggest: given the filtering rule, we would only expect 0.35 spurious POPs in the F0 column. Given that expectation, it is certainly possible that one ($p = 0.25$) or maybe even two ($p = 0.05$) false POPs could have crept in, but very unlikely that false POPs make up an appreciable proportion of the total of 45.

Table 3: Number of *usable* pairwise comparisons, by #loci and #excluding loci. Comparisons are *not usable* if the adult was caught in or before the year of juvenile birth, and/or the false-positive probability was too high (see text). Columns 8-21 omitted for brevity.

.	F0	F1	F2	F3	F4	F5	F6	F7	>	F22	F23	F24	F25	TOTAL
C11	.	.	.	.	1	4	5	21	>	.	.	.	.	84
C12	.	.	5	42	340	1345	4019	9114	>	.	.	.	.	57,000
C13	.	1	16	151	887	3420	9900	20482	>	.	.	.	.	143,000
C14	1	4	61	587	2876	11277	32947	70962	>	.	.	.	.	652,000
C15	.	3	42	375	1962	8411	27165	66386	>	.	.	.	.	923,000
C16	2	1	18	131	966	4716	17097	47526	>	.	.	.	.	1,170,000
C17	2	.	8	92	655	3674	14677	45482	>	.	.	.	.	1,942,000
C18	5	.	6	65	483	2699	12037	40524	>	.	.	.	.	3,063,000
C19	7	.	1	33	288	1728	7992	29511	>	.	.	.	.	4,158,000
C20	2	1	1	15	131	886	4630	18722	>	.	.	.	.	5,512,000
C21	14	.	1	5	62	481	2589	11387	>	.	.	.	.	7,197,000
C22	.	.	.	.	4	38	165	698	>	117	.	.	.	1,170,000
C23	4	.	.	.	2	20	143	754	>	2383	179	.	.	2,966,000
C24	2	.	.	.	4	22	90	558	>	17376	2799	214	.	5,097,000
C25	6	.	1	.	1	5	22	199	>	42419	10339	1607	139	4,123,000
SUM	45													38,180,182

It is also possible to compute an "expected" version of Table 3, assuming there are no true POPs. That is: for each comparison, taking into account which loci were used, we can compute the probability that there were 0, 1, 2, ... mismatching loci if the pair was truly unrelated. By summing the probability of, say, 1 mismatching loci over all comparisons with, say, 11 loci, we can compute the expected value of the (C11, F1) element corresponding to Table 3. The left-hand columns of the result are shown in Table 4, after filtering out the same comparisons as in Table 3. By definition, the row-totals would be the same as in Table 3; the question is how close the column totals are, as shown in the bottom two rows of Table 4. And they are very close, except of course for the F0 column where we are seeing true POPs. This is good; the laws of probability seem to be working well today. The close correspondence between observed and expected totals for F1/F2/F3 suggests that the calculations leading to 0.35 expected false POPs are sound; of course, the *actual* number cannot be exactly 0.35, but it is most likely 0, and most unlikely to be more than 2.

Table 4: *Expected* number of comparisons with a given number of mismatching loci, given the loci actually used in each comparison, and assuming no true POPs. The TOT OBS row at the bottom is taken from Table 3.

	F0	F1	F2	F3
C11	.	.	0.02	0.17
C12	0.02	0.63	9.43	82.46
C13	0.04	1.27	17.91	149.85
C14	0.15	4.50	60.38	491.97
C15	0.08	2.56	36.35	315.94
C16	0.03	0.98	15.15	144.87
C17	0.02	0.55	9.16	94.78
C18	0.01	0.30	5.32	58.88
C19	.	0.15	2.75	32.39
C20	.	0.05	1.12	14.42
C21	.	0.02	0.47	6.48
C22	.	.	0.02	0.30
C23	.	.	0.02	0.25
C24	.	.	0.01	0.13
C25	.	.	.	0.03
TOT EXP	0.35	11	158	1392
TOT OBS	45	10	160	1496

Using a cutoff to exclude ambiguous comparisons does entail a bias-variance trade-off, because some true POPs may have been overlooked in the excluded comparisons, and any reduction in the overall number of POPs found will increase the uncertainty in our final estimates. However, given the threshold we used, it is only when the number of loci compared is 14 or less that substantial numbers of comparisons are excluded (from comparison of Figure 2 and Table 3), and overall only about 5% of comparisons are excluded. Thus we have managed to achieve less than a 1% bias while only incurring a $\sqrt{5} \approx 2\%$ increase in standard error compared to what we would have gotten from "perfect" genotyping (where every pairwise comparison is usable). This reflects very well on the tissue quality, the processsing, and the selection of powerful, reliable loci.

Of the 45 POPs found, it is interesting that 9 included one locus where the two animals were scored as different homozygotes (one AA and the other BB). We had deliberately relaxed the exclusion rule to permit this situation, in case of "heritable nulls" (see Appendix, section 6.3), and there was no ambiguity about the POP status of these pairs based on the remaining loci[6]. In all but one of the 9 cases the apparent mismatch occurred in one or other of the two loci which exhibited substantial excess homozygosity (D569 and D573; see Appendix), consistent with the "heritable null" possibility.

Note also that close-kin relationships at the level of uncles-and-nephews, while possibly as common as POPs in reality, are not going to lead to false POPs in this study. Between an uncle & nephew, only 50% of loci will share an allele by descent anyway, so with these loci the overall chance of sharing an allele is about $1/2 * 1 + 1/2 * (1 - 0.65) = 0.68$ (compared to about 0.35 for an unrelated pair), and the chance of getting say 20 loci all sharing an allele through chance is about 0.0004— so there would need to be about 2000 uncle-nephew-level pairs to generate a single false POP.

[6]Including additional checks at the extra one or two loci which were not normally used in mass-screening for POPs

2.1.1 Cases where no POPs should be found

As an exercise, we can repeat Table 3 just comparing juveniles with themselves, where true POPs are impossible; see Table 5. The expected total in the F0 column is again 0.35; this time, the observed total is 1 (in C13/F0, so towards the lower end of the number of loci compared) which as noted earlier has about a 25% probability and gives no indication that the false-positive calculations are failing.

We can also compare all adults with all other adults (not shown). This time, POPs are actually possible, albeit likely rare because of the time required to reach maturity— see later discussion. There is indeed one possible POP (C18/F0; unlikely to be by chance, given 18 loci used), and it is plausible biologically. The female "parent" was aged 24 when caught in 2007, and the female "offspring" was 177cm (not aged, but any age from 12 up is plausible, given other length-at-age data) when caught in 2009; this gives plenty of scope for the "parent" to have been mature when the offspring was born.

Table 5: Comparison of juveniles to themselves.

	F0	F1	F2	F3	F4	F5	F6	F7
C11	.	.	.	.	.	.	6	9
C12	.	.	8	45	329	1404	4611	10109
C13	1	.	7	63	399	1574	4935	10697
C14	.	1	36	257	1335	5386	15948	35522
C15	.	1	15	153	872	3307	10661	25493
C16	.	1	6	42	304	1465	5341	14986
C17	.	.	2	31	232	1236	4744	14436
C18	.	.	2	26	169	1010	4318	14160
C19	.	.	6	21	144	888	4136	14761
C20	.	.	1	14	85	603	3025	12153
C21	.	.	1	.	37	275	1644	7109
C22	.	.	.	.	.	22	97	524
C23	.	.	.	.	6	14	98	524
C24	.	.	.	1	2	8	69	403
C25	.	.	1	.	1	6	23	115

2.2 False negatives?

What about accidentally excluding true POPs? That can only happen if there is genotyping error[7]. Large-scale errors involving multiple loci at once would be (and were) detected and fixed by our QC procedures described in the Appendix, so the concern here is about small-scale errors at a single locus and specimen. If such errors lead often to false-negative POPs, these should show up low down in the F1 column of Table 3, as near-POPs that apparently fail to match at one locus (false-negatives at multiple loci being correspondingly rarer). That is not what is seen; rows C17 down have only one entry in F1, compared with an expected total of 1.1 from Table 4.

Prior to producing Table 3, we independently re-scored[8] all the apparent true POPs in F0, all the F1s, and F2s in the rows from C17 down. The original version of Table 3 had 44 rather than 45 POPs; the re-scoring moved one pair from C15/F1 to C17/F0 (changing one existing

[7] Or mutation, but with say ~50 POPs and ~20 loci each, and mutation rates thought to be about 10^{-4} per generation, mutation is unlikely to have happened amongst our POPs.

[8] "Re-scored" means: we re-examined all the peaks and came up with new scores, but did not re-do any of the chemistry.

score, and scoring 2 more loci originally deemed unscorable). The lower left-hand corner of the Table (apart from true POPs in F0) was still empty even without rescoring. Although rescoring changed only about 1 POP, it does give some indication of scoring error rates. Across the 1400 loci that were rescored, there were 8 individual changes, plus deleting one panel of loci for one fish; four of the changes were to delete a score altogether when a locus looked dubious, and the other four were to add a second allele to a "homozygote" (a definite error). Note that all 8-9 changes in the rescoring only unearthed one false-negative (corrected in Table 3), so the *effective* false-negative rate for POP purposes seems to be well under 0.5%. It would also be possible to produce per-locus estimates of scoring error rate based on the partial re-runs and re-used control fish in our QC procedures.

The most important line of evidence to suggest that false negatives from individual scoring errors are not a serious problem, though, remains the absence of entries in the lower left-hand corner of Table 3. Appendix 2 presents a formal statistical approach to estimating false-negative rates by comparing Tables 3 and 4; the point estimate of the overall number of remaining false-negatives is in the range 1-2, and the upper 95% CI in the range 2-3. In any event, false negatives must be at most a small proportion of the 45 POPs.

2.3 Summary of genetic results

Extensive QC procedures were used to ensure consistent and reliable scoring throughout the project. In all, we conducted about 40,000,000 pairwise comparisons to look for POPs. A few pairs had to be excluded because they had too few scored loci to reliably screen out unrelated pseudo-POPs. However, because of the number and quality of loci used, we were able to choose a cutoff for exclusion that implies very little bias (i.e. unlikely to unearth false POPs) while incurring very little penalty in variance (i.e. using nearly all the comparisons). QC protocols were devised to catch large-scale mixups. With respect to small-scale (individual-level) scoring errors, the error rate is too low to cause a substantial proportion of true POPs to be overlooked. In all, we found 45 POPs in about 38,000,000 usable comparisons.

3 Qualititative findings about the POPs

3.1 Sex, age and size of parents vs general adults

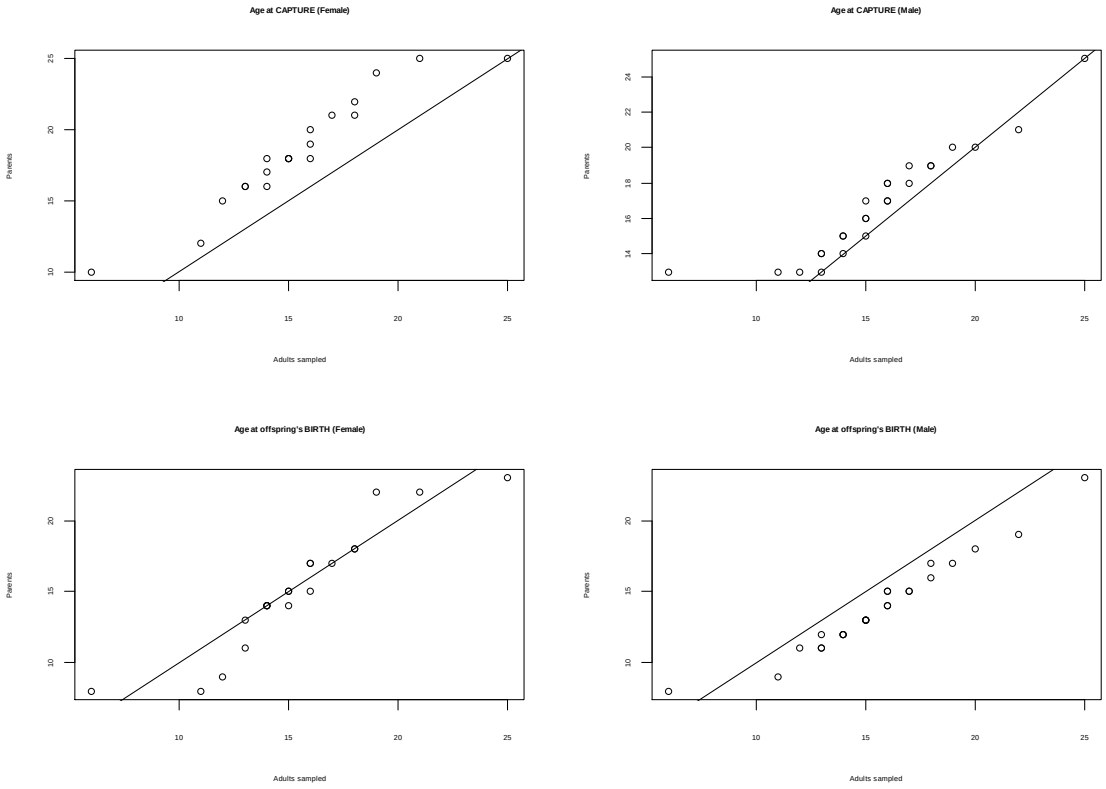
Of the 45 POPs, 20 were female and 25 male. All adults in POPs have now been aged; about 1/3 were aged under the Indonesian/Australian ageing program, and the remainder were aged specifically for this project after being identified through genotyping. On average, parents *at capture* are somewhat older (and bigger; not shown) than typical captured adults of the same sex. However, this comparison is not "fair" because the parents have had the opportunity to grow during the interval between juvenile birth and adult capture, which in this study is on average about $3\frac{1}{2}$ years.

Since juvenile age is known (3 in almost all cases), it is easy to back-calculate parental age when the offspring was born. The youngest successful spawners were aged 8, for both sexes[9]. When back-calculated parental age is used instead of age of capture, the difference between parental and typical adult age distribution disappears for females, and actually reverses for males; But it is important to realize that this back-calculated comparison is also not "fair". Adults are subject to selectivity bias in favour of bigger/older fish, and the selectivity pattern

[9] An earlier version of this document reported an anomalous 6-year-old spawner. On re-checking the otoliths, it turned out that the two otoliths for this fish actually came from different adults, so there must have been a handling error at the collection site. Based on the correct otolith (i.e. the one whose DNA matched the main adult tissue sample), the actual spawning age turned out to be 9.

on the parents would also have changed over the interval between giving birth and being caught. Back-calculated age distributions will be skewed towards younger/smaller fish, compared to what would have been found if the same set of parents had somehow been sampled in the year of offspring birth.

Figure 3.1: QQ plots of parental age vs adult age, by sex . Points left/above the line mean parents are bigger/older. Females on left, males on right. Upper graphs show parental age at year of capture, lower graphs at year of birth.



The upshot of this rather involved argument is:

- parents at capture are older/bigger than typical adults, because they have aged/grown since giving birth;
- back-calculated parental age distribution at offspring birth is similar to typical adult age, but...
- the back-calculated distribution is biassed towards smaller/younger fish, so...
- female parents would actually be bigger than typical adults if it was possible to sample them in the birth-year.
- It's not clear whether the same would be true for males.

These phenomena can only be fully disentangled with the aid of a mini-assessment model.

Similar results are found using length rather than age, but the mini-assessment then has to be fitted beforehand, because of the need to back-calculate length.

3.2 Skip-spawning

From the small number of POPs identified in time for CCSBT 2011, there was no obvious indication of skip-spawning. However, the larger sample of POPs now available does show evidence of biennial spawning for younger fish. The test is to take each POP, and note how many years actually elapsed between juvenile birth and adult recapture, vs how many years *could* have elapsed given the POP was eventually found. For example, if the juvenile in a POP was born in 2007, then only comparisons with 2008/2009/2010 adults would be meaningful, so the probability of matching to a 2008 adult is roughly[10] equal to the proportion of adults checked in 2008 relative to those checked in 2008+2009+2010. Table 6 shows the results, split by parental age at offspring's birth; for younger parents, almost all observed gaps are even-numbered, but not for older parents. The pattern is not sex-specific.

Table 6: Distribution of gap between Juvenile-Birth-Year and Adult-Capture-Year, for young & old parents. Dot means zero. Right-hand table is condensed to odd/even gaps.

Age	Gap (years) ->	1	2	3	4	5	6	7	Age	Gap->	Even	Odd
8-12	Obs	1	6	.	2	.	4	.	8-12	Obs	12	1
	Exp	1.6	2.3	2.7	2.6	1.9	1.1	0.6		Exp	6.1	6.9
13-25	Obs	7	5	10	7	2	1	.	13-25	Obs	13	19
	Exp	4.3	6.8	7.4	6.2	4.5	1.8	1.0		Exp	14.9	17.1

Any errors in ageing would obscure patterns such as seen here. Although the sample size is not huge, the difference for younger adults is significant at 1%.

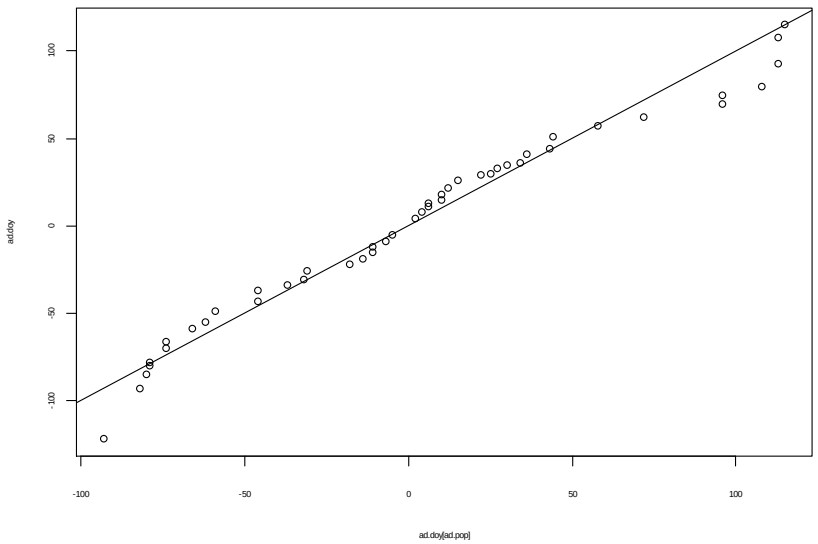
Skip-spawning is not a particular problem for this close-kin study because the study covers many years and the even/odd effect should largely wash out; the general effect of smaller fish being less present is already allowed for in the mini-assessment, because average spawning-ground residence-time (including the probability of not being on the spawning grounds at all) gets estimated as a function of length and sex. However, in a more perfect world, probabilistic size/age-based skip spawning would be allowed for in the mini-assessment.

3.3 Timing in spawning season

Parents of GAB juveniles have the same distribution of capture date within season as do "average adults" (Figure 3.2). Thus there is no evidence of "temporal stock structure" in a way that might lead the abundance estimates to be biassed (eg we *might* have seen that parents of GAB juveniles always spawn early, and we might *not* have had equal coverage through the Indonesian fishing season). Breaking down by sex does not reveal anything either.

[10] Calculations are approximate: e.g. the "expected" rows do not account for growth or mortality, but should reflect any even/odd pattern OK.

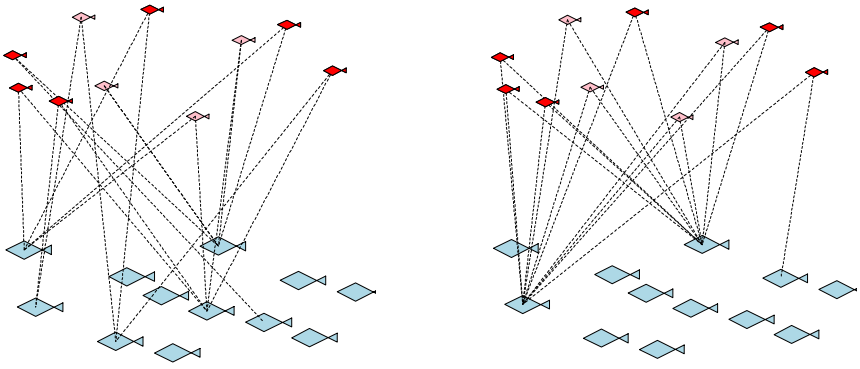
Figure 3.2: QQ plot of day-of-year of capture of Parents (X) vs Adults-in-general (Y)



3.4 Incidence of (half-)siblings among the POP juveniles

There are none. In other words, none of the POP adults match to more than one juvenile. That is a good thing, because if (half-)sibs are common among the *sampled* juveniles, then the pairwise comparisons become non-independent. Figure 3.3 shows what might happen; if there are many (half-)sibs in the juvenile sample, then the number of links to parents remains the same so the abundance estimate is still unbiassed (noting that an adult can "count" in more than one POP), but its variance would increase because the number of POPs actually found would depend critically on whether the "super-parents" were caught.

Figure 3.3: Cartoon depicting the impact that reproductive variability would have. Small fish are juveniles, red ones are sampled.



A preliminary check in 2010 just among juveniles indicated that (half-)sibs could not be *very* common (a critical decision point for the project), and the 7 POPs found in 2010 contained no sibs or half-sibs. Having found none in this much larger set of POPs, we can maybe conclude that (half-)sibs are rare enough *among our juvenile samples* for their effects on variance to be ignored. This is not to say that (half-)sibs are at all rare among *all* 3-year-olds, but simply that our juvenile samples are a very small fraction of the total, and are well-enough-mixed to

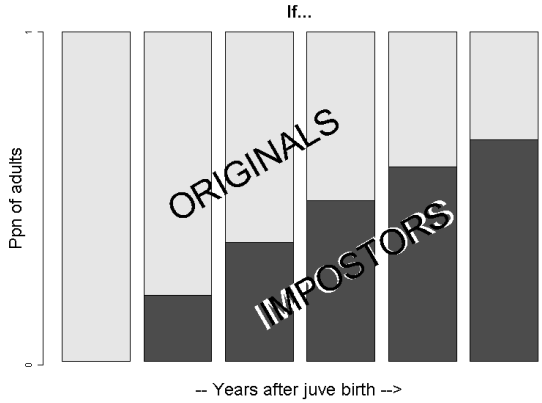
make sib-pairs rare. As an academic exercise, it will at some point be interesting to re-run the juvenile-only sib check with the greatly expanded set of loci now available.

4 Mini-assessment

4.1 What effects need to be considered?

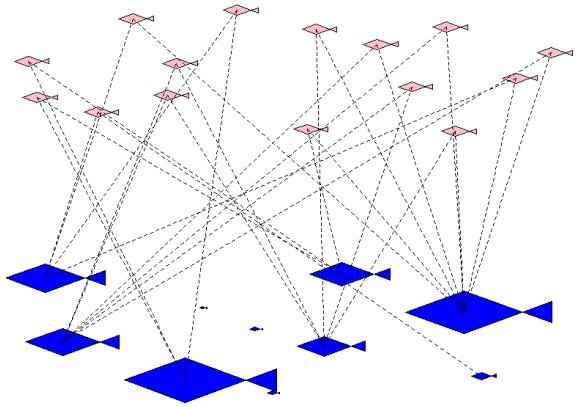
There are two main reasons why the $2m^2/P$ "cartoon" estimator ("cartoon" in the sense of Figure 3.3, for example) would be seriously misleading for SBT. The first is that we cannot do comparisons only against the "parental cohort-group" of each offspring, i.e. the group of adults that were alive at its birth. Figure 4.1 illustrates the main point; if survival rates are the same for all adults, then the cartoon estimator would still be valid even with time lags, *provided* we could restrict comparisons to the light-grey parental cohort-group. But we cannot do so, because (i) we do not know the age of all adults sampled, (ii) maturity is not knife-edge so there is no absolute definition of the parental cohort-group, and (iii) maturity is quite likely length- rather than age-driven. If we are forced instead to sample adults from say the entire 4th column of Figure 4.1, after a 3-year gap, then a high proportion of comparisons will be with "impostor" adults that could not have been parents, and the cartoon estimator would be biassed high.

Figure 4.1: Dilution of original parent-cohort-group by incoming recruitment



The second, linked, reason is that adult sampling is strongly selective towards large/old fish, which are also likely to have been more fecund (even allowing for a 3-year time lag). Because they are more fecund, they have more "tags" per capita (i.e. juveniles that they are parents of), and each tag is more likely to be "recaptured" (i.e. the adult is more likely to be caught) because of selectivity in favour of larger adults. This is the close-kin analogue of "heterogeneity in capture probability", a well-known issue in mark-recapture abundance estimation. Figure 4.2 shows the cartoon version. The upshot for the naive $2m^2/P$ estimator would be that each comparison is more likely to yield a POP than would a comparison with a randomly-chosen adult.

Figure 4.2: Big tuna are more fecund *and* more likely to be caught



Both effects concern not the number of POPs actually found, but rather the difficulty of working out how many comparisons are "relevant" for abundance estimation. The two effects act in opposite directions; the time-lag dilution means that some comparisons are invalid and thus less effective than "random" comparisons, whereas the selectivity-fecundity correlation means that the valid comparisons are more effective than "random" comparisons. The time-lag dilution is also mitigated by growth, since the surviving "original" adults after 3-4 years will be bigger than the "impostors" and thus more likely to be caught. However, there is no particular reason to assume the effects will cancel out, since the time-lag effect is driven primarily by the length of the study whereas the selectivity-fecundity effect is determined by the nature of the fishery and the growth curve. And the effects can be quite large; with an adult survival of say 0.8, after 3 years only 50% of the original adults are still alive to be sampled and the impostors will (in equilibrium, and neglecting selectivity and growth) be involved in about 50% of the comparisons. To deal with these issues properly, we need a mini-assessment.

One further issue arises from of the extended timespan of this study, which spans juvenile birth-years from 2002 and adult capture-years to 2010, as well as the initial age structure of the adults in 2002 which was determined by even earlier events. The 1990s and 2000s have been eventful decades for SBT, and it may be such that steady-state assumptions are simply not viable.

A proper close-kin abundance estimate for SBT therefore has to deal with survival, selectivity, fecundity, and growth, and perhaps with changes in abundance over time. The requisite data come from the length and age-at-length samples from Indonesia, plus fecundity studies explained below. While not strictly "fishery-independent", length and age data are not subject to the same problems as CPUE or total catch. It also makes sense to split the analysis by sex: the cartoon applies equally well if applied to males and females separately, where the chance of a POP comparing to a male adult is $1/N_{\text{male}}$ not $2/\left(N_{\text{male}}+N_{\text{female}}\right)$, and C is split into $\left(C_{\text{male}}, C_{\text{female}}\right)$.

4.1.1 Residence time, selectivity, and fecundity

The tropical waters off Indonesia are really no place for an adult SBT, an animal that is adapted superbly for much cooler temperate waters. Adults arrive on the spawning grounds fat, and leave thin. Of course, the longer they can stay on the grounds, the more chances to spawn they will have, so it seems reasonable to suppose that they will put up with Indonesian conditions for as long as their bodies let them. The key for disentangling the effects of fecundity, survival, and selectivity, is average **residence time** on the spawning grounds, as a function of length. A cursory glance at length distributions from Indonesia shows that few fish under 150cm, and none under 130cm, are caught on the spawning grounds, so there is obviously some link to

length. As per the skip-spawning discussion, "average residence time" already factors in the probability that a fish won't be there at all in any given year. Our mini-assessment specifically assumes that, given length and sex:

- Selectivity $\propto$ residence time
- Annual reproductive output $\propto$ residence time $\times$ daily reproductive output

Except as specifically noted later, we assume that length and sex are the driving influences behind the behaviour of adult SBT, rather than age.

Of course, there could be other "second-order" phenomena which slightly change the above relationships (e.g. different depth distributions by size, and thus different exposure to hooks; different egg *quality* with parental size; etc etc) but these seem likely to be small compared to the dominant effect of residence time. For the rest of this document, it may be helpful to think about selectivity and residence time as directly equivalent, at least within each sex.

We have no direct data on residence time as a function of length, so the relationship needs to be estimated indirectly from data. Independent data on residence time and depth distribution as a function of length, from archival tags placed on big fish, would be extremely useful: both in tightening up parameter estimates in our existing model, and in assessing whether the effects that we hope are "second-order" really are.

4.1.2 Fecundity analyses: daily reproductive output

The canonical reference for SBT (female) spawning biology and fecundity is a study from the early 2000s by Davis et al.[11]. In summary, female SBT while on the spawning grounds have an on-off cycle, consisting of several days of consecutive daily spawning (one spawning event per 24 hours), followed by several days of rest while more eggs are built up. This on-off cycle may be repeated several times. As soon as the final spawning cycle is complete, they leave. The mass of eggs released per daily spawning event can be estimated from the change in gonad weight between just-about-to-spawn and just-after-spawning fish; it is approximately proportional to length to the power 2.47. The average duration of each part of the cycle (and thus the proportion of days on the spawning grounds when spawning actually occurs) can also be estimated as a function of body length using histological data, because the first day of a spawning sequence can be distinguished from the other days, and similarly for a resting sequence. However, the number of cycles per season is completely unknown, and is obviously set by the residence time.

To summarize, the factors involved in daily reproductive output are:

- reduction in gonad weight per spawning event
- duration of consecutive spawning day sequences
- duration of consecutive resting day sequences

A reasonable amount of data is available for all three of these, and the relationship to length can be estimated from fitting three GLMs. (This was already done for the first two factors in Davis et al., and the third factor was addressed during this study.) For now, we have treated the parameter estimates as exact in the rest of the mini-assessment.

We have no comparable data for males, nor on the extent to which male abundance actually influences the number of fertilized eggs per year.

[11] T. Davis, J. Farley, M. Bravington, R. Andamari (2003): *Size at first maturity and recruitment into egg production of southern bluefin tuna* FRDC project 1999:106

4.2 Indonesian length, sex, and age data

A substantial proportion of the Indonesian SBT catch is sampled as it passes through the main landing port of Benoa. Length (to the centimetre) and sex are always recorded, and nowadays otoliths are always extracted, although only a length-stratified subset (500 per year in the recent past) are read. Between 900 and 1700 animals were measured per year between 2002 and 2010. Thus the data can be seen as

1. Random samples of length and sex from the entire adult catch
2. Random samples of age, given length and sex.

Even without the POP data, it is possible to do some steady-state analysis of the age/length/sex data (though it is obviously impossible to estimate absolute abundance), but it is impossible to completely separate selectivity (as a function of length) from average adult survival rate. When the survival rate is very high (e.g. 0.9) or very low (e.g. 0.5) it does become impossible to match the observed length-frequency distributions except by invoking a ludicrous selectivity function, but in the absence of other data reasonable fits to the age and length data can be obtained across a wide range of survival rates.

Fortunately, the POPs can help estimate survival rate, in addition to absolute abundance. The typical gap between offspring birth and adult capture— assuming that the adult is in fact captured subsequently, i.e. that the pair is an identified POP— is related to survival. If survival rates are low, very few parents will survive to be caught say 7 years later (the maximum gap possible in this study), so most of the POPs that are found will be separated by just one or two years. Growth and residence time need to be properly accounted for too, but the intuitive basis should be clear. The close-kin data thus has three vital roles: the *number* of POPs (given the number of comparison) essentially sets the scaling of absolute abundance, the age and length distribution *within* the POPs informs on selectivity/fecundity, and the distribution of time-gaps *within* the POPs essentially determines survival.

4.3 Model structure

The model keeps track of numbers by age and sex; each year, each fish either gets one year older or dies. However, most phenomena are driven by length, which is assumed to have a fixed distribution at age. Each fish has its own personal L_∞, drawn from a sex-dependent log-Normal prior whose mean and variance at age are fixed, while the other von Bertalanffy parameters are the same within each sex. A plus-group is used for ages 25 up, and a minimum "recruitment" age for possible spawning also needs to be set (currently 8). There is also a plus-group for length (200cm) and, unusually for stock assessments, a sort of "minus-group" as well, currently set to 150cm. Experience with fitting just to age and length data showed that trying to extend the fit to the small proportion of adults below 150cm gave poor results, in that this small "tail" started to "wag the dog" and distort the fit elsewhere. The focus of this study is spawners, which are mostly 160cm and up, so it is more important to get a good fit there than to squeeze a last drop of misinformation out of very small adults. However, it is necessary to somehow keep track of the small spawning contribution of fish in the minus-group, and accordingly there is some tedious book-keeping code in the model.

Most of the likelihood is quite standard; multinomial distributions for length-sex frequency data, and for age given length and sex. The effective sample sizes of the length and age data were capped at 300 per year, to avoid these data swamping the information from the POPs. The novel term is the contribution of the POPs. For each comparison made between a juvenile j and an adult i of sex (gender) g_i, the outcome (POP or not) is a Bernoulli random variable

with probability given by

$$\mathbb{P}[j \sim i] = \frac{\text{expected ARO from } i \text{ in year of } j\text{'s birth}}{\text{total ARO from adults of sex } g_i \text{ in that year}}$$

where ARO is Annual Reproductive Output, i.e. daily fecundity multiplied by residence time as in section 4.1.1. This formula replaces the "$2/N$" probability in the simplest possible close-kin implementation.

To actually compute a likelihood, it is necessary to specify various terms:

- numbers-at-age in 2002, and for incoming recruitment (age 8) in 2003-2010;
- survival rate in each year and age;
- residence/length relationship;
- growth parameters;
- relation between daily RO and length *for males*.

The total number of potential parameters is colossal because of the numbers-at-age and survival terms, so of course one needs to specify them parsimoniously given the limited amount of data available. This is done using formulas (sensu R) for each bullet-point term, describing what covariates are allowed to influence it, and perhaps what functional form that influence might take. For example, we might choose to make survival constant over age and time, except for the plus-group[12]. We might also make assumptions of constant "recruitment" (at age 8) in the 2000s; and/or that numbers-at-age prior to 2000 were in equilibrium with survival; and/or that von Bertalanffy k is the same for both sexes; and/or that the slope of the residence/length relationship (but not its midpoint) is the same by sex; etc. One example is given in the Results section.

The final term— male daily reproductive output as a function of length— can *in principle* be estimated provided we are willing to assume that survival rates for males are the same as for females. Without that assumption, there is nothing to anchor the selectivity/survival/fecundity triangle for males. For females, we do not need to estimate this term because we have direct data from the fecundity studies.

The likelihood itself is coded in Pascal, with derivatives computed by an automatic differentiation toolbox similar to ADMB. The overall data-handling and fitting is done in R, calling the `nlminb()` optimizer to do the fitting. Some care was needed to avoid numerical problems in calculating the log-likelihood, and because of limited time there are still starting-value problems so that some model parametrizations can't yet be fitted. However, once a starting value has been obtained, no convergence problems were encountered, at least for the fairly parsimonious specifications (say 15 parameters) that have been tried to date.

4.4 Results

It will be apparent that an enormous number of different *versions* of the mini-assessment could be run. A full investigation is far beyond the scope of this project, and should probably be undertaken in conjuntion with other data sources, particularly to provide context on recruitment levels. A limited set of versions has been run, but the results actually shown here come from an almost-steady-state version of the model, with constant adult survival and constant recruitment

[12] In SBT as with other top-predators, it must be the case that natural mortality rate increases for old animals, since simple maths shows that the sea would otherwise just fill up with decrepit tuna.

from 2002 onwards but an age composition in 2002 that need not correspond to a steady-state prior to 2002.

Basic investigations across suggest that:

- Mean L_∞ is appreciably larger for males than females. The evidence for any difference in k or t_0 is not overwhelming, but making these two sex-linked as well does not seem to overparametrise the model. CV of length-at-age appears to be the same for both sexes.

- Residence time appears to be lower for males of a given length than for females, so we do need a sex-specific intercept in this term. However, there is not enough data to estimate any sex difference in the *slope* of the relationship. Also, introducing extra flexibility in model form beyond the logistic (asymptotic) can give nonsensical predictions for very large fish. A good choice seems to be `~sex+length`.

- There is not much information for estimating male daily reproductive output as a function of body length. We have assumed instead that male daily output is directly proportional to length (i.e. exponent of 1). There is no good reason for that particular choice, but fortunately the abundance and survival estimates seem not to be much affected by assumptions about male daily output in practice, even though it could matter in theory.

- Based on just one comparison: changing the annual effective sample size for length/age data from 300 to 900 did not much affect the abundance estimates much (i.e. by a few percent).

Having got these basic issues out of the way, the remaining questions are how to set up initial numbers-at-age and incoming recruitments. In a full steady-state model, the age distribution in year 1 (actually 2002AD in our setup) is determined by the survival rate, and the incoming recruitments thereafter are equal to the numbers at recruitment age in year 1.

Some diagnostic plots for the steady-state model are shown in Figures 4.3-4.4. These pertain to the length and sex data only, since the POP data are really too sparse for diagnostics. The length-frequency data, shown for few years only in Figure 4.3, are mostly not too bad despite the steady-state assumption, except for 2002 where the data seem completely different from other years. The fits to age-at-length are very good (not shown). However, there is a problem with the fits to sex ratio by length class (Figure 4.4): in the biggest length classes lower down the graph, where males tend to predominate thanks to their bigger asymptotic size, there is a strong decrease in proportion of females over the 2000s. This decrease is seen overall too (in the black dots), but is not apparent in the smaller lengths, where there is a rise followed by a dip. This difference in trend across length classes suggests that methodological changes in how sex is assessed are unlikely to be the cause. The underlying cause needs some further thought; it has nothing much to do with close-kin, and is a question for the OM as well as this mini-assessment.

Figure 4.3: Steady-state diagnostics: length. The unreadable parts do not need to be read; they show details of the particular model version fitted.

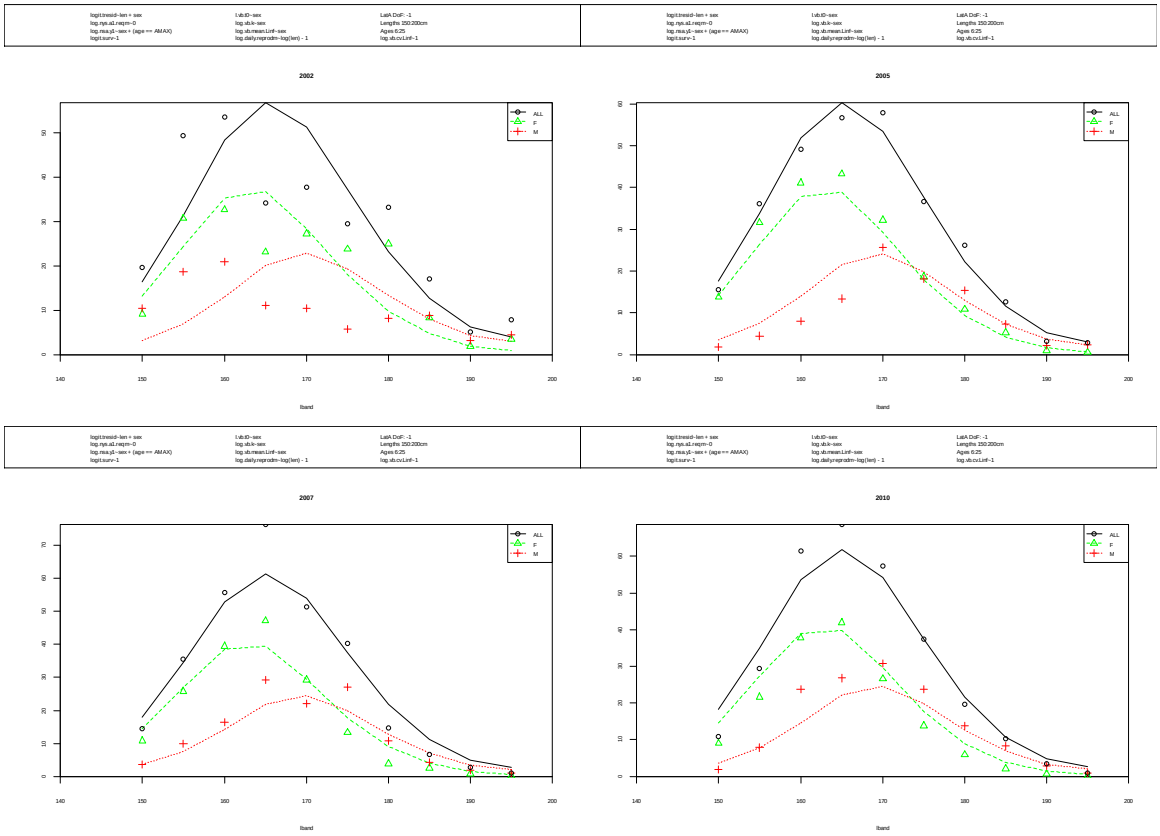
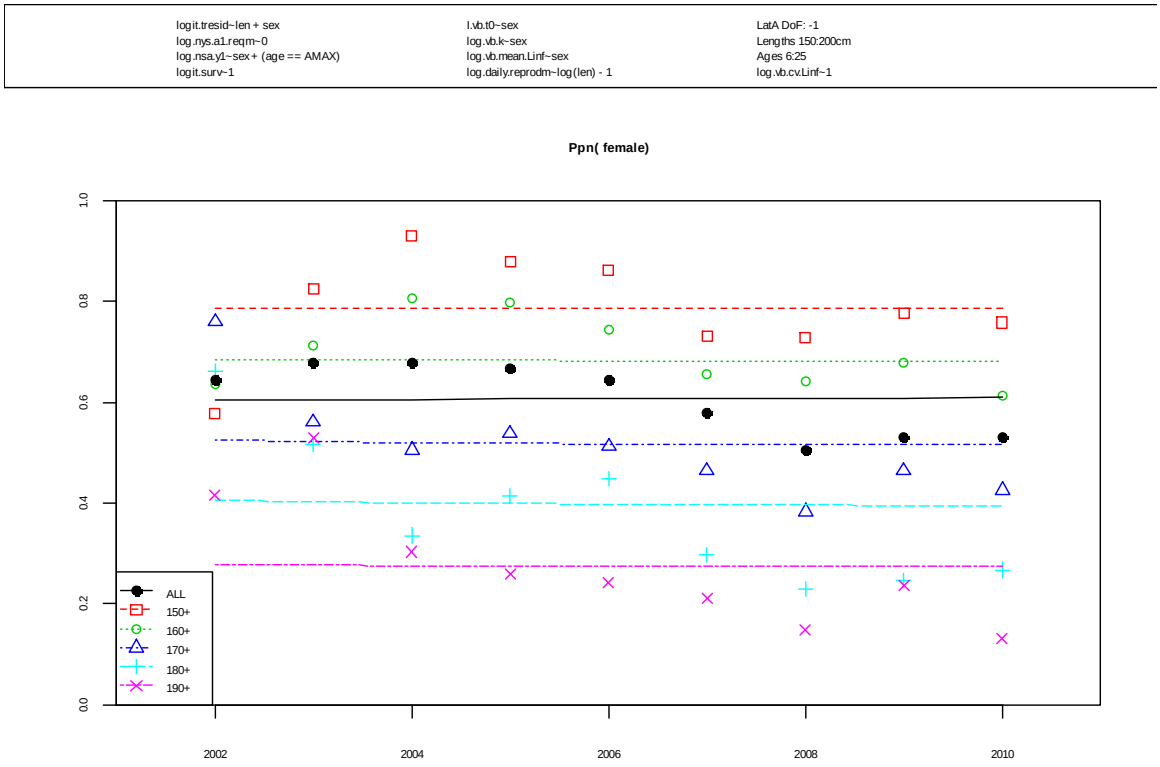


Figure 4.4: Steady-state diagnostics: sex-ratio. The unreadable parts do not need to be read; they show details of the particular model version fitted.

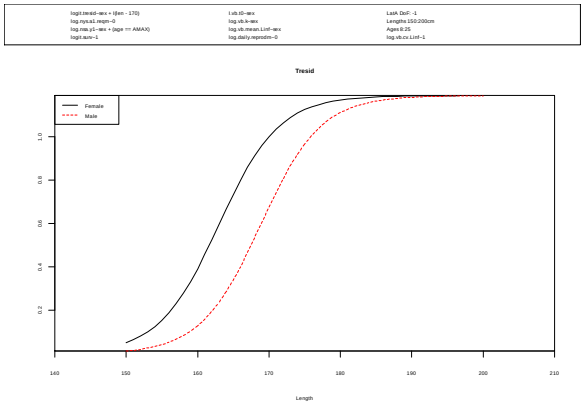


4.4.1 Parameter estimates and uncertainty

Annual adult survival for the steady-state model was estimated at 0.73, fairly close to OM estimates. However, the estimated abundance of 10+ adults in 2004 is much higher than in the OM: 2.04M fish, with a biomass of 157kT. This happens to be fairly close to the simple "twice the comparisons divided by the POPs" estimator, but only by coincidence; the competing effects of dilution by incomers, growth, and selectivity are all strong, and merely happen to largely cancel each other out.

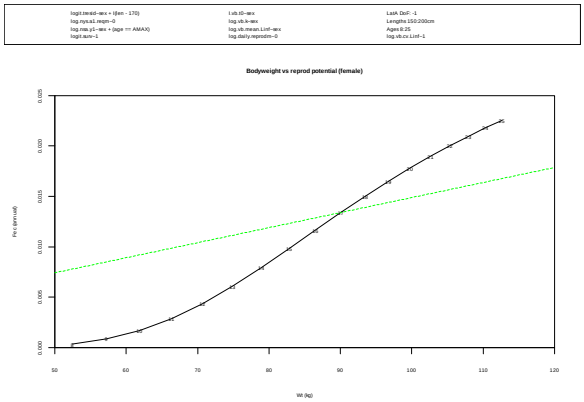
The estimated relationship between residence time (i.e. selectivity, in this model) and length is shown in Figure 4.5; the curve climbs steeply from 160cm for males and about 155cm for females, with males taking longer to "mature". The apparent asymptotic slowdown around 180cm may be a consequence of the functional form chosen (a logistic curve), and warrants further study.

Figure 4.5: Residence time as a function of length



By combining the estimated residence-time with the estimated growth curves (which have average L_∞ of 191cm for females, and 201cm for males) and the fecundity data, it is possible to infer the average female spawning contribution at age. The results are very different to what is assumed in the existing OM, i.e. that spawning contribution is proportional to biomass for ages 10 and up (Figure 4.6); the mini-assessment suggests that older fish are comparatively much more effective spawners than younger fish. This underlines the point that the mini-assessment is structurally different to the OM, and results need to be compared with that in mind.

Figure 4.6: Relative spawning contribution as a function of female bodyweight. Average bodyweight at ages are indicated. Green line corresponds to current OM assumption.



The nominal CV of $\hat{N}\,(10+, 2004)$, obtained by inverting the Hessian, is 17.3%. This is only slightly higher than the theoretical lower limit of 14.9% set by the intrinsic sampling noise associated with the observed total of 45 POPs (see Appendix 3 for why the POP count is unlikely to be overdispersed). The concurrent estimation of survival, residence time, etc., as well as abundance, has contributed only an additional 2.3% to this nominal CV. However, the CV is still "nominal" because that additional contribution is not finalised yet; for one thing, it depends on the "effective sample size" used for the length frequency data, which we simply assumed to be 300 fish per year. Given that there is substantial variability (noise) between observed and predicted length- and sex-compositions, the additional CV should probably be higher than 2.3%. Even so, in an exploratory run with the effective sample size for length frequencies set to just 30 fish per year, the additional CV still only reached 5%, i.e. a total CV around 20%.

To improve and fully de-nominalize the CV, it will be necessary to handle better the overdispersion/variability in the non-POP data, so that the "likelihood" in the mini-assessment accurately reflects the uncertainty. Also, experiments with non-steady state versions of the mini-

assessment indicate that the model uncertainty associated with possible past & present *trends* in recruitment is substantial (though less so around 2004 than in the most recent years). Moving to a random-effects rather than fixed-effect framework for recruitment should mitigate this, and help to accommodate the model uncertainty automatically in the final result. These changes will push the CV up somewhat but, from what we have seen so far, we would be surprised if the final CV[13] exceeds 25%.

Accordingly, we plan to make just a few further changes to the mini-assessment model in 2012, in the course of submitting for peer-reviewed publication and finalizing the CSIRO/FRDC project report. The main ones are:

- allow recruitment (annual incoming 8-yr-olds) to be a random-effect, rather than constant or a trend;
- formally estimate the "effective sample size" for the length-frequency data;
- formally propagate the uncertainty associated with the fecundity GLMs.

With our current mini-assessment, these issues would have to be treated as "model uncertainty", which we have not attempted to explore in this report. Preliminary explorations suggest that (i) the first two can have an appreciable effect on the point estimates of abundance, though by no means enough to change the qualitative conclusions, and (ii) the CV is probably not going to increase much. Once we have tackled these issues inside the mini-assessment, the associated uncertainty should be reflected directly in the final CV.

5 Discussion

This project has successfully managed to complete an enormous amount of genotyping with tight quality control. The data do seem to be able to deliver an internally-consistent fishery-independent[14] estimate of adult abundance, just as planned. The stand-alone estimates are still somewhat preliminary, but are clearly considerably higher than the OM estimates— at least 3 times the point estimate from the "base case" scenario, and on the edge of the upper confidence interval of the most optimistic scenario. While this may seem surprising, it should be emphasized that there is very little reliable data in the OM with which to estimate absolute adult abundance (although other quantities such as relative depletion can be estimated more reliably), which of course is in itself is a big part of the reason for undertaking this study. Also, as noted in section 4.4.1, the different notions of effective fecundity in the two models make direct comparisons tricky. So getting an adult abundance estimate that is very different to the OM's is by no means an indication of a serious problem with either the CK estimate or the main conclusions of the OM. Nevertheless, it is important to ask the obvious question: how wrong could these CK estimates be? There are a limited number of issues to consider, given that we are not asking about small changes here— the point is to try to think of any phenomena that could make a huge reduction to the estimate, of the order of 50%.

5.1 Is the number of POPs about right?

The genetic results strongly suggest that there are few if any false negatives or false positives, given the filtering we have used, the proven effectiveness of our large-scale QC checks in detecting and fixing problems, and the absence of "near-misses" in the lower left-hand corner of

[13] It is important to choose a sensible quantity to estimate the CV of. For example, once recruitment is allowed to vary annually as planned for our final model, the average 10+ abundance across 2002-2008 will have a lower CV than the corresponding abundance in any single year.

[14] Strictly: catch and CPUE independent

Table 3. An independent implementation of the QC software checks would be useful, particularly when the data migrate to our existing SBT database, and this conceivably might unearth a few further problems. However, while it is certainly possible that there are a small number of false negative or false positives in (or not in) our POPs, there is strong evidence that the proportion is not substantial.

5.2 How precise is the estimate?

The nominal CV of our steady-state example estimate is 17.3%, driven mostly (15%) by sampling variability in the number of POPs found. That 15% component seems solid, since there is no reason to expect substantial overdispersion in the number of POPs; see Appendix 3 (section 8). The effective sample sizes we assumed for the length data seem to have been somewhat high, so the current additional contribution of 2.3% from the length/sex/age aspects is probably too low but not by all that much, going by our explorations so far. The remaining changes planned for the mini-assessment will include uncertainty related to those effective sample sizes, and will accommodate aspects of model uncertainty not currently captured in the nominal CV. Based on our explorations so far, though, we will be quite surprised if the final CV exceeds 25%. And, apart from the proviso about selectivity discussed next, the final CV should be a statistically defensible measure of overall uncertainty.

5.3 Is the abundance estimate about right, given the number of POPs?

If the number of POPs is about right, and if the adult sampling is simultaneous with juvenile birth and random, then the cartoon estimate can't go wrong— each juvenile really does have exactly two parents. Most of the other potential problems with close-kin— stock structure, or massive proportions of sibs/halfsibs— don't apply to SBT. So the only other source of possible error is in the adult-assessment model. As mentioned at the end of section 4.4.1, the model is not completely finished and the remaining modifications will change the point estimates somewhat, but we do not expect those changes to be very large. So, aside from possible programming mistakes (this is still a very recent assessment, all coded by one person), there are two main points to consider:

1. The entire CK assessment, and the way in which the cartoon adjustments are implicitly calculated, rests on the assumption that selectivity is primarily driven by residence time— the longer a fish is on the spawning grounds, the more likely it is to be caught, all else being equal. The link between residence time and annual female reproductive output rests on the same assumption (more spawning opportunities). It is hard to see how these assumptions could actually be wrong, but the caution might be in the phrase "all else being equal". If there are other really major length-based effects on selectivity or on reproductive output (aside from female daily fecundity, for which we at least have some data), then bias could perhaps arise.

2. The only other way that an abundance estimate could be biassed, is if there is some type of heterogeneity between adults that is *not* just due to length and sex, and which results in some adults (i) being more likely to spawn offspring caught in the GAB, *and* (ii) more likely to be caught themselves in the Indonesian fishery at least one year later. It is hard to imagine what might cause such heterogeneity.

Even if there do turn out to be errors in these estimates, they seem more likely to be programming errors, and therefore fixable, rather than being intrinsic problems with the data or

its interpretation. The CK data fundamentally do seem to be extremely useful for SBT: they are bearing out their promise. There is obvious scope for continuing to collect and genotype in future, both to build up the time series and also (thanks to the retrospective qualities of close-kin) to increase the number of POPs found from our already-genotyped juveniles from 2006-2010. The way this might fit into SBT management, and the links with other monitoring possibilities, is far more than can be explored in this study, but the potential value of further CK genotyping is clear.

Finally, we draw attention to the key role of residence time on the spawning grounds— or, to be accurate, how the average residence time depends on size— in getting to an actual abundance estimate, and a selectivity estimate, and an appropriate definition of spawning potential. Although there is just about enough data in the POPs and the age/length samples to infer the residence/size link indirectly, it would be immensely useful to have direct estimates from a few adult fish across different sizes, since this could both ground-truth the model and give a basis for estimating further length-dependent effects on selectivity, if that turned out to be necessary. Pop-up satellite tags could yield limited information quickly, but the best data would come from archival tags because they can record over several years, and are not as vulnerable to short-term tagging shock. The low fishing mortality on adults means that quite a few archival tags would be needed to get recaptures, and that we might have to wait a while to get the tags back, but the number of returned tags needed would not have to be at large (even single figures) to give a very useful check on, and input to, close-kin based abundance estimates in future. Such tagging ought not be a very expensive exercise in terms of the value of the fisheries, or indeed the cost of this close-kin project.

6 Appendix 1: The genotyping and QC process

6.1 Terminology

This section is meant as a guide for a non-geneticist in a CCSBT context. It is not intended as an authoritative set of definitions from a genetic perspective, which are widely available in the genetics literature. These definitions include forward-cross-references in *italics*, and use **bold** to indicate additional definitions. In the text after this section, a few technical genetic terms have been used and marked with an asterisk, but deliberately not defined since their relevance will only be apparent to those who already understand them.

Locus: an identifiable place on the genome with characteristic start and end sequences of DNA, and a variable DNA sequence between them. The loci we used are **diploid,** so that each individual has two versions (**copies**), one copy being inherited from each parent. The sequences of the two copies might be different or might by chance be the same. We used **microsatellite** loci, whereby each sequence is characterized simply by its length or **size** (i.e. the number of *nucleotide bases* it contains), which will be a integer in the range say 80-600 depending on the locus and how it is to be purified away from the rest of the genome in any particular study.

Alleles: the set of possible sequences a locus can have, i.e. for microsatellites a set of integers. Alleles at different loci might happen to have the same length, but are in no sense comparable— it only makes sense to refer to an allele for a specific locus. The **allele frequency** for the locus is the frequency distribution of the different alleles across the population under study. A **highly variable** locus has a large number of different alleles and an allele frequency that is not dominated by just one or two common alleles. The probability that two unrelated animals will have an allele in common is lowest if the locus is highly variable, so such loci are preferred for close-kin work. A **null** allele is an allele

that is present in the animal, but is not revealed by genotyping; possible causes include scoring error, and a mutation in or near the locus that causes the DNA amplification process to fail for that copy.

Genotype: which alleles an animal has. Usually means for all the loci together (sometimes called a **multilocus genotype** or **DNA fingerprint**), but can mean just the alleles carried at a single locus if specified.

Homozygote/heterozygote: An animal is said to be a **heterozygote** at some locus if the two copies are different alleles, or a **homozygote** if they are the same.

Scoring/genotyping/calling: deciding which alleles are present at a locus for a particular animal. This really involves many steps, but sometimes "scoring" just refers to the final step of adjudicating on the possible alleles proposed by the *GeneMapper* software. The protocol in our study is that, if the genotype at a particular locus for a particular specimen is ambiguous, no score is recorded (rather than trying to make a subjective best-guess).

Scoring error: Recording the wrong genotype at one or more loci. Large-scale scoring errors affecting many fish and loci simultaneously can arise from inadvertently swapping or rotating entire plates of fish, or from miscalibration of the *sequencer* for a particular *run plate*. Small-scale scoring errors affecting individuals most commonly involve failure to detect a small second peak in a heterozygote, so that the locus is mistakenly scored as a homozygote instead. At least in this study, actual mislocation of peaks were very rare (based on a subset of the fish which were independently re-scored).

Amplification/PCR: the chemical process by which the DNA from certain desired loci only is selected and amplified for input to the sequencer.

(Nucleotide) base is one genetic "letter" (C/G/A/T), the molecular building-blocks which are linked together to form a DNA molecule. DNA occurs in two strands, and each base is paired with its complement on the other strand, so the term **base pair** is often used instead.

Tetranucleotide: The sequences within microsatellite loci are mostly repeats of some short subsequence of base pairs, such as GATA (four base pairs, so a tetranucleotide locus) or CA (a dinucleotide). Dinucleotide loci are more common in most genetic studies, but are more prone to scoring error. In this study we used only tetranucleotide loci.

Panel refers to a set of loci (usually 4-7) which can be analyzed simultaneously by the sequencer.

Plate is a group of 96 DNA samples (including a couple of controls— standard specimens included on every plate— and blanks) placed in wells numbered A1-H12 in an industry-standard format on a small rectangular tray ("**96-well microtitre plate**") ready to load into a sequencer. Each group of 96 fish is originally set up on a **template plate** from which are prepared several **run plates**, all with the same layout of specimens in the 96 wells, but with each run plate specific to a particular panel of loci.

Sequencers are the machines that physically do the genotyping. One run plate is **run** or **sequenced** at a time. For each locus on each specimen, the output is a graph with X-axis corresponding to allele length (as a continuous variable) and "signal intensity" on the Y-axis. Alleles are visible as peaks with a characteristic shape.

GeneMapper is software which identifies possible alleles from the sequencer's output. In most cases, GeneMapper will propose the correct peaks, but each sequencer graph and proposed scores is scrutinized by an experienced scorer who makes the final decision on which peaks truly represent alleles, and which peaks are artefacts.

Bins and binsets: Because of slight variations in run conditions, the locations of peaks reported by GeneMapper will vary fractionally between sequencer runs, even for the same sample. **Bins** are therefore used in GeneMapper to provide tolerance and to convert the continuous-valued peak locations into an integer-valued allele size. Each bin is a continuous-valued range such as $[137.2, 138.6]$, which should span the range of peak locations found for that allele across many runs. The **binset** for each locus is the collection of all its bins. The binset needs to be consistent throughout a study. In this project, we initially developed bins and binsets from genotyping the first 500 individuals, then revisited them after 5000 specimens had been genotyped.

6.2 Flowchart of genotyping

For the last 9000 of the 14000 fish genotyped (from both sites), the procedure was as follows.

1. Tissue biopsy samples from each fish are collected, labelled, and stored in boxes of 100 fish, with corresponding information on fish length, date of capture etc, and in the case of the Indonesian fish are cross-linked to the existing data (sex, otolith if collected, etc).

2. The original tissue from each fish is subsampled down to the 10mg size suitable for DNA extraction. The remaining biopsy tissue is archived, so this (relatively costly) step can be repeated if necessary.

3. The DNA of 96 subsamples at a time is extracted into solution. Part is kept frozen as an *archive plate*. Part is used as the template plate, incorporating two controls (in specified positions) and two water blanks (in known positions, variable from plate to plate).

4. For each template plate and panel of loci, a small amount of fluid is used to prepare a run plate. Enzymes are used to amplify the desired loci (for the panel) from the rest of the genome, and their DNA is PCRed in a multiplex(*) reaction at CSIRO.

    a) The first column of each run plate is then duplicated as a column on a *check plate*, of which there is one per 12 run plates.

5. Run plates are sent to AGRF[15] in Adelaide for sequencing. There are four run plates per 96 fish, labelled H/I/J/L depending on panel of loci are involved. For each run plate, the result is a set of 96 "FSA files" suitable for input to GeneMapper.

6. FSA files are scored at CSIRO by an experienced team; we have used only four scorers, each of whom scored several thousand samples. Results from the various scorers have been cross-checked for consistency on some plates.

    a) The check plates are sequenced at CSIRO using similar machinery to AGRF's, and the results are compared to the corresponding columns of the FSA files from AGRF. This provides a safeguard against plates being swapped or rotated, and against miscalibration of the sequencer.

[15] Australian Genome Research Facility

b) The panels all include a common locus B8B, so by comparing the B8B scores across run plates ostensibly from the same template plate, we could check whether the files for each run plate really did come from their nominal samples.

7. A variety of QC checks are run on the FSA files, to detect plate-level phenomena such as rotation/swapping/miscalibration (see 6.1), atypical allele frequencies, and excess homozygotes, and individual-level phenomena such as duplicate genomes which arise if samples are inadvertently double-sampled at the point of collection.

For the first 5000 fish genotyped, a slightly different and less streamlined procedure was used in steps 4-5. Only 20 of the 27 final loci were used. The PCR for some of the 20 was done by AGRF in single-plex(*) reactions which were then grouped into three panels A/B/C to make run plates, while the remaining loci were multiplexed at CSIRO as above into two run plates corresponding to a pair of panels D & E. [The A/B/C loci were subsequently reorganized into two of the panels (I & J) used for the last 9000 fish; the D/E panels became the H & L panels after the 7 extra loci were added.] After the FSA files returned to CSIRO, we used a shared locus on the D & E panels to check their "alignment", as in 6b. To check alignment of the A/B/C panels with each other and with the D/E panels, we put DNA drawn from the first column of the template plates for D/E panels into one column of an extra template plate, which was then used to make run plates for the I & J panels (containing the same loci as A/B/C, but organized differently). These were sequenced, scored, and the genotypes compared against the corresponding columns in the original A/B/C plates. Although this process was somewhat cumbersome and led to some duplication in scoring (about 10%), it provided an important safeguard against the handling errors that become almost inevitable with such large sample sizes.

6.3 Selection of loci

Loci for this project went through an particularly extensive checking process. Past experience indicated that, to be conservative and to facilitate automated genotyping, we needed to strictly focus on using tetranucleotide repeats that gave solitary, sharp, allele peaks. In short, we wanted a set of loci that: were highly variable but not so variable that the longest alleles failed to amplify well; had simple peak structure with minimal shoulder to the peaks and little stutter; and had clear gaps between alleles. Over time, as more fish were scored, some of our best tetranucleotide loci turned out to have some two-base-pair insertion/deletions, which meant that some alleles were separated by only two base pairs (though usually at least one of the alleles involved was rare). This was tolerated, provided there was at least a one-base-pair gap between bins. Loci were immediately discarded during the initial testing phase if they showed alleles separated by just one based pair, indicative of poly-nucleotide tracks in the amplified allele. After genotyping 5000 fish, we had developed 20 loci organized into 5 panels A-E, with very comprehensive scoring binsets into which almost all detected alleles fell. At this point we included an additional 7 loci (total 27) which were re-organized into four multiplex panels H, I, J, and L. We scored all 27 loci where possible in the remaining 9000 fish, but used only 25 loci for finding POPs; the remaining two loci, with slightly less reliable scoring, were used only for QC purposes, as per 6b above. When scoring, our protocol was not to record a score if in doubt, which is safe for purposes of POP-finding.

An important check in genetic studies, is on the proportion of homozygotes found at each locus. In theory, provided a number of assumptions hold, this can be predicted from the allele frequencies, and the extent to which there is an excess of *apparent* homozygotes is one indication of the reliability of a locus. As shown in Table 7, all but 3 of the 25 primary loci have both low expected homozygosity (which corresponds to being highly variable, and thus powerful for POP

identification), and at most a small excess observed homozygosity as given by the difference between the EXP and OBS rows; this suggests relatively few cases of failure to see the 2nd peak in a heterozygote, for example[16]. The exceptions are in the bottom right of table: D569 and D573. It appears (as described shortly before section 2.1.1) that the excess of homozygotes in those two loci is due to "heritable nulls" (eg from a mutation in the flanking sequence(*) so that primers(*) don't bind), so that some alleles simply don't amplify. No loci showed appreciable evidence of Short-Allele Dominance(*).

To guard against the possibility of heritable nulls in *any* locus, a comparison of two different apparent homozygotes (AA in one fish vs BB in the other) was *not* used to exclude a POP even though there is ostensibly no shared allele, in case the real score was "A-null vs B-null" with the null being inherited. This relaxation has only a small effect on the false-positive probability. However, it is not feasible to relax the exclusion criterion further to allow for the commonest (but still fairly rare) scoring error whereby the second peak of a heterozygote is missed, i.e. by treating AA vs BC as not necessarily excluding. Such a weakened criterion would generate many false positives with the existing set of loci, so many more loci and more expense would be required.

Table 7: Homozygote percentages, "expected" (ignoring nulls) and observed; see text.

	3D4	B5	D10	D111	D11B	D12	D122	D201	D203	D211	D225	D235	D3	D4D6
EXP	19.8	6.8	7.1	11.8	10.7	10.8	9.7	11.7	7.5	11.4	3.4	8.5	16.8	5.5
OBS	19.8	7.3	7.3	12.2	12.3	10.9	11.4	12.4	9.0	17.0	3.7	14.8	16.8	6.7

	D541	D524	D549	D570	D592A	Z3C11A	D517	D534	D582	D569	D573
EXP	14.0	12.4	11.9	7.3	9.8	13.0	3.1	9.3	7.6	9.9	4.9
OBS	14.0	13.5	11.9	7.3	10.2	13.4	3.4	10.1	7.7	45.5	30.9

6.4 QC for Consistency of Allele Size Calling

Examining the consistency of allele-size calling is fairly straightforward, and is mostly dealt with by use of an internal standard and use of an automated genotyping program developed by ABI-Life Technologies (supplier of the DNA sequencer used for fragment separation). To further minimise inter-run variation, all size fragmentations were run on only one DNA sequencer located at the Australian Genomic Research Facility (Adelaide node). This eliminated variation occasionally observed when the same samples are run at two facilities even on the same model of sequencer.

In addition, the ABI system uses an internal size standard added to each sample from which the size curve is extrapolated for estimating allele peak length relative to the standard curve. ABI states that variation using this system ensures +/- 0.5bp accuracy from run to run. Furthermore, the GeneMapper program analyses each individual size curve for peak quality and general fit to the theoretical ideal size curve. Any discrepancies detected by the software raise flags in the analysis window and can be scrutinized in further detail. We also examined each size curve analysis as well as the individual peaks that were used to generate the size curve for each individual in a run plate to ensure another level of QC in addition to that used by the GeneMapper software.

GeneMapper uses a standard set of allele size bins used to smooth out further subtle variation and ensured easy comparison among alleles from different individuals and provided another

[16] Other reasons for deviation from Hardy-Weinberg, such as population structure, are unlikely for SBT, and in any case no deviation is seen for many of the loci despite the very high sample sizes and consequent high power to detect any deviation.

level of QC among plates. Bin sets are developed for each locus to permit automated genotyping using the GeneMapper software. Individual bins represent a value range centred on the median length value of each allele as ascertained following sizing of an initial set of individuals. Preliminary bin sets were developed following detailed analysis of about 500 fish. These sets were designed to encompass slight variations to permit detection of gross deviations from the norm greater than +/- 1.0 bp. After genotyping about 5000 fish, the bin sets were re-assessed for consistent allele calls, and a final consensus adjustment was determined. Bins permit assignment of an integer value to the continuous-valued allele length based on the GENESCAN size standard, and permit simple comparison of allele identities among individual genotypes. A gap of one to three base pairs between bins ensures that an objective decision rule can be consistently applied to a genotype for inclusion of an allele into a designated integer bin. Alleles falling in the gap were rare and presumed to be a result of an insertion or deletion event on an individual's DNA. These were scored as "unknown genotype" but the real value could still be used for confirmation of parentage should it be required to confirm identity (not required with our samples to date).

The use of automated genotyping with a single set of GeneMapper bin-sets allowed us to detect if peaks were consistently falling outside of predetermined bins and would highlight a general problem with the running of a plate (eg. old buffer or polymer in the sequencer leading to general failure of proper electrophoresis and inconsistent separation). Runs where problems were found were re-run with new buffer and polymer; this rectified the problems in every case.

6.5 Avoidance of chimeras

Chimeric genotypes are (in this study) a composition of DNA from more than one fish, rather than (as in some other studies) DNA profiles resulting from multiple DNA in a well (two or more contaminated DNA leading to more than two alleles present for each locus). There are only two possible sources. First, a chimeric error will result from turning a run plate 180 degrees, whereby e.g. the A1 position became the H12 position. This error produces what looks like a legitimate DNA profile but made up of some loci from fish A1 mixed with the remainder of loci from H12 from the run plates that were not rotated. Second, if two run plates are swapped, the loci for those panels (but not for the other panels on the same fish) will be swapped. Clearly, these errors will lead to any POP members on the plate being overlooked, affecting 100-200 fish at a time, so it is important to catch them. Fortunately, once one is aware of these possibilities, it is fairly easy to write QC software using the check-plate results and/or the controls to detect and fix the problem. We did find both types of chimera in this study (rarely), but thanks to the QC protocols we were able to detect and fix them.

6.5.1 Further processing details for the first 5000 fish

For the first 5000 fish we developed a unique system to cope with the potential issues arising from PCR and fragment separation methods used at the outsourcing facility (AGRF). The first 5000 fish were run at AGRF as three single-plex (A, B, and C) and two multiplex (D and E) panels. At this point the multiplex PCR was clearly the most optimal solution and we included 7 additional loci that were incorporated into an optimised set of four multiplexed panels (H, I, J, and L). The A, B, and C panels were combined into the I and J panels while D and E were combined into H and L. To check for generation of chimeric genotypes we used the set of template plates that were the source of DNA for the D and E panels. Since D and E had a common locus scored for both plates we were able to ensure that there were no chimeric individuals there. We then ran the first column of each template plate for panel-I and for panel-J. This checked the genotype calls of 8 individuals that should be identical if no mix up had occurred. We verified that all 8 genotypes for each locus was congruent across all tested

plates indicating that no single-plex mixups had occurred. Since the template plates used were those used to set up D and E we were then assured that there were no chimeric fish generated in the first 5000 genotyped individuals.

6.5.2 Further processing details for the last 9000 fish

For the balance of the fish, a unique system to identify individual template plates was developed to ensure that the fish on the plate could be identified, and that it was not accidentally rotated prior to sequencing. The four panels had a common locus to check on plate to plate variation, and also to detect PCR contamination via negative water controls. Template plates were created in a specific routine fashion with four positions in each plate reserved for positive and negative controls. We used two positive control individuals on every plate with position A01 being control fish #1 (TC-2005, male) and G12 being control fish #2 (TC-2205, female). The positions of the negative water controls were used to uniquely identify each plate. For example, one plate would have water controls in position A02 and A07, while the next plate would have A02 and A09. Care was taken to ensure that the water was placed in one odd-numbered and one even-numbered well row due to the way the 48 capillary sequencer picked up the samples; every dip of the sequencer thereby had one positive and one negative control, so that each electrophoresis had internal controls to check run quality. The internal common locus control for each individual checked to see that each fish was scored with consistent fragment separation for each of the panels. By use of this system for the final 9000 fish, we were able to QC for chimeric individuals, check for PCR contaminants in the master mix, ensure that run conditions did not affect genotype scoring among the four panels, and also ensure that plates were not mislabelled or loaded into the sequencer incorrectly. Our QC caught a few errors but these were few and subsequently dealt with by a quick rerun of the PCR or fragment separation or both.

7 Appendix 2: Rigorous estimation of false-negative (FN) rates

The question of interest is: what proportion of true POPs could have a scoring error that leads to the POP being overlooked? We can estimate this directly by comparing Table 3— observed numbers of (loci compared, loci failing to match)— with Table 4 (expected version of Table 3, assuming zero POPs and therefore zero FNs). If the expected-value calculations behind Table 4 are correct, and if there are numerous true POPs without FNs, then Table 3 should resemble Table 4 except for numerous entries in the F0 column— which is pretty much the case. If the Table 4 calculations were wrong for some reason[17], then the upper-right-hand triangle of numbers in Table 3 would be stretched to the left compared to Table 4— which is not the case. Therefore, we can take the expected values in Table 4 as correct if there were no POPs, and use the differences between the tables to make inferences about the true number of POPs, and about how many FNs are in Table 3. We can do this because FNs will appear in Table 3 as an "echo" of the F0 column, predominantly in column F1, and somewhat weighted towards the lower rows because there is more chance of a scoring error when more loci are involved. Apart from chimeras and mass failures of PCR on a run plate, as described and ruled out in Appendix 1, there seems no reason why scoring errors should not be independent across loci

[17] The only theoretical reason we can see why the calculations in Table 4 might ever go wrong, is if genotypes at different loci within each fish are not independent, something which could arise from substantial cryptic stock structure, with different allele frequencies in the different stocks. That situation is *a priori* unlikely for SBT, and happily there is no suggestion of it in Table 3.

on the same fish; hence, provided scoring errors are uncommon to begin with, FNs are most likely to be in the F1 column, less likely to be in F2, and rapidly less likely beyond that.

The numbers in Table 3 actually result from a second round of checking; we re-scored all the pairs in the F0 and F1 column, and in the lower rows of the F2 column. However, only a small percentage of the fish were re-scored during the second round, and the level of attention paid to these fish may not be typical of the rest of the sample. In this section, we have therefore analysed the data from the *preliminary* version of Table 3, before any selection of fish to re-score took place. This makes the analysis general, but also means that the results are pessimistic in terms of FN likely FNs compared to the final data, because the FN/near-FP status of many would have been cleaned up during re-scoring. The preliminary data, shown in 8, is very similar to Table 3, the main difference being that the C23 row starts (3,1) rather than (4,0); this is one case where a scoring error did cause a false-negative, though this was subsequently detected and fixed on re-scoring. The other differences did not affect POP status of any pairs.

Table 8: **Preliminary** number of *usable* pairwise comparisons, by #loci and #excluding loci, **before** re-scoring. First three columns only.

.	F0	F1	F2
C11	.	.	.
C12	.	.	5
C13	.	2	16
C14	1	4	61
C15	.	3	42
C16	1	1	18
C17	3	.	7
C18	5	.	7
C19	7	.	1
C20	2	1	1
C21	14	.	1
C22	.	.	.
C23	3	1	.
C24	2	.	.
C25	6	.	1
SUM	44		.

7.1 Likelihood for estimating false-negative rate

Let θ be the probability that a pair of fish will be a POP (so θ is inversely related to abundance, etc), and let e be the probability that one shared locus in a POP will fail the parent-offspring compatibility test[18], either through mis-scoring or mutation. Assuming scoring errors at different loci are independent[19] and equally likely[20], then the probability of f loci failing in a POP where c loci are compared, is a simple Binomial probability. Also, for a non-POP pair where c loci are being compared, let p_{cf}^{NON} be the probability that f of the loci will fail the test. For

[18]The basic test is: do they share a visible allele? We used a more relaxed version, so that AA vs BB homozygotes are also deemed (potentially) compatible.

[19]Apart from chimeras, as described and ruled out in Appendix 1, and mass failures of PCR on a run plate which would be picked up by our other QC checks, there seems no reason why independence could fail.

[20]Strictly, the probability of a scoring error that leads to rejection of POPhood probably varies somewhat across loci, but there is not nearly enough data to estimate this; and since the set of loci that actually get used in a comparison is a random variable, and we are only concerned with one or two errors here, the approximation is statistically negligible.

any given pair, this actually depends on the particular loci involved, and is already calculated to form the basis for the expected values in Table 4. Any given pair with c loci compared is either a POP or not, and the probability p_{cf} that the pair will fail at f loci is therefore

$$p_{cf} = \theta \left(\begin{array}{c} c \\ f \end{array} \right) e^f \left(1-e\right)^{c-f} + \left(1-\theta\right) p_{cf}^{\mathrm{NON}}$$

Therefore, if n_c denotes the number of comparisons using c loci in Table 3, the expected value of cell (c, f) is $n_c p_{cf}$. Strictly, the distribution within each row is Multinomial, but in the first few columns the multinomial "size" is enormous (millions) and p_{cf} is small, so a Poisson approximation is perfectly adequate. If y_{cf} denotes the observed number of pairs in the (c, f) entry of Table 3, then the likelihood of the first few columns up to F failures is (up to a constant)

$$\prod_{c=11}^{25} \prod_{f=0}^{F} e^{-n_{cf} p_{cf}} \left(n_{cf} p_{cf}\right)^{y_{cf}}$$

The term p_{cf} involves the parameters θ and e, which can be estimated via maximum likelihood.

The bulk of the information on false-negative rates is contained in the F1 column (and the F0 column, which is needed for estimating θ), with a little coming from the F2 column. To the right, the noise from the increasingly large numbers of almost-false-positives swamps any signal related to false-negatives with 2, 3, etc number of failures, which will be increasingly rare.

7.2 Confidence intervals on actual FNs

Although the Hessian from the above likelihood could be used in the standard way to derive a confidence interval for the *expected* number of FNs in a *replicate* of this study, that would be solving the wrong problem. Our interest lies in the *actual* number in *this* study; so, if FNs were very unlikely beyond the F1 column, then the number of FNs would be capped above by the total F1s seen, regardless of how many might be found if the study was repeated. This makes quite a difference in practice. A Bayesian argument is required to get the answer we need.

We need the probability distribution of the number of false-negatives $\#FN$ given the observed data, i.e. $\mathbb{P}\left[\#FN | y\right]$ where $\#FN$ is the total number of False Negatives and $y = (y_{cf} : c \in 11 \cdots 25,\ f \in 0 \cdots 1)$ is the observed numbers in the F0 and F1 and possibly F2 columns (F3 onward are irrelevant because the chances of 3 or more scoring errors is negligible). For simplicity of argument, say for now that we neglect the F2 column as well. Obviously, the maximum possible value of $\#FN$ is the observed number of F1s, in this case 12. Each of these F1 pairs is either a near-FP or an FN. The probability that an F1 pair with c loci compared is actually a FN rather than a near-FP, is

$$\frac{\mathbb{P}\left[\text{1 error in } c \text{ loci}\right] \times \mathbb{P}\left[\text{is POP}\right]}{\mathbb{P}\left[\text{1 error in } c \text{ loci}\right] \times \mathbb{P}\left[\text{is POP}\right] + \mathbb{P}\left[\text{match at } c-1 \text{ of } c \text{ loci}\right] \times \mathbb{P}\left[\text{is not POP}\right]}$$

One implication is that a (C12,F1) fish is much more likely to be a near-FP than a (C25,F1) is, because (i) the probability of a non-POP matching by chance at 11 of 12 loci is much higher than for 24 of 25, and (ii) the chance of a scoring error is about twice as high with 25 loci as with 12.

The FN-status of the pairs are independent[21], θ and e, so the total number of F1 pairs that are FNs is the sum of (in this case) 12 independent Bernoulli (0/1) random variables, with

[21] I.E. the probability that a given F1 pair is actually FN or near-FP is unaffected by the FN-status of the other F1 pairs, given θ and e.

probabilities depending on the number of loci involved. There is an algorithm for calculating the Bernoulli-sum probability distribution, which is already used in the expected-FP calculations[22]. Hence, given a pair of values (θ^*, e^*), we can easily compute $\mathbb{P}[\#FN = x|y, \theta^*, e^*]$ for $x \in 0 \cdots 12$. What we actually need, though, is

$$\mathbb{P}[\#FN = x|y] = \int \mathbb{P}[\#FN = x|\theta, e, y]\, f(\theta, e|y)\, d(\theta, e)$$

which can be estimated by repeatedly drawing pairs (θ^{*j}, e^{*j}) from the posterior distribution of $(\theta, e|y)$ via importance-sampling, and then averaging the $\mathbb{P}[\#FN = x|y, \theta^{*j}, e^{*j}]$ across all the draws. This requires a prior for (θ, e), which we took to be independent uniform on $\log\theta$ and logit$e$, plus of course the likelihood from section 7.1. A fully-conditioned confidence interval on $\#FN|y$ can then be found simply by inverting the cumulative distribution of $\#FN|y$.

7.3 Results of FN analysis

We ran the above algorithms first on just the F0 & F1 columns of Table 8, and then on the F0, F1, and F2 columns. In the first version, the Maximum Likelihood Estimate on #FNs was 1.95 and the 95% UCI was 2.46; in the second version, the numbers were 3.19 and 4.0. The difference is entirely driven by the (C25, F2) entry, discussed further below; without it, the two versions are almost identical. Both versions indicated a very low *expected* number of FNs in the F2 column or beyond (less than 10% of the number expected in F1), although the second version clearly identified an *observed* likely-FN at (C25, F2).

As noted above, these FN estimates are *prior to* rescoring the F0, F1, and F2 (from C16 down) columns. Rescoring certainly fixed one FN, at (C23, F1), so the appropriate estimates and limits for the number of FNs in our final dataset (after re-scoring) are no more than (MLE 0.95, UCI 1.46) or (MLE 2.19, UCI 3.0).

The nature of the mismatching loci for any pair provides additional information on whether an F1 or F2 pair is really a FN, as opposed to just being a lucky near-FP from an unrelated pair. This is because one type of mismatch arises from a comparatively common scoring error (overlooking one allele, so a fish is recorded as AA when it should be AB), whereas the other type (incorrect size for an allele) is extremely unlikely; this was apparent in the results from our routine QC rescoring exercises of individual fish. In particular, after carefully rescoring the (C25, F2) pair, the only way it could be a FN POP would be to have a mutation at one locus and a scoring error at a second— a very unlikely conjunction of events. However, this pair is also a very unlikely event under the only two other possible scenarios: an exceptionally-matched unrelated pair, or a well-matched uncle-nephew-pair (which must be much, much rarer than unrelated pairs). In the end, the only way to resolve the true status of the (C25, F2) pair will be to use more loci, which we plan to do as part of a different project. We cannot at present decide whether to treat (C25, F2) as a FN (in which case we should use the second version of the FN analysis, including the F2 column, to get a point estimate of about 2 FN), or not (in which case we should use the first version, with a point estimate of about 1 FN).

Thus, further detailed investigation of the rescored F1s and F2s might eventually shed some light on whether we should expect 0, 1, or 2 FNs in addition to our 45 POPs. However, whichever the answer, the analysis in this Appendix demonstrates that the proportion of FNs to true POPs must be small, and is certainly not going to affect the qualitative conclusions of this project.

[22] K Butler, M Stephens (1993): The distribution of a sum of Binomial random variables. Tech Rep 467, Department of Statistics, Stanford University

8 Appendix 3: What might cause overdispersion in the POPs?

The CV of the "cartoon" abundance estimate is just the CV of the number of POPs found. We have treated this as "count data", so that its variance is equal to its mean. The question arises: under what circumstances might there be overdispersion in this count?

Overdispersion would arise when the 38,000,000 comparisons are substantially non-independent. It's easy to see why a high frequency of (half)sibs would do that: if every juve had one full-sib partner in the sample, then the results for one sibling completely predict the results for the other, and the information content would only be that of 19,000,000 independent comparisons. (Recall that each POP is counted, even if the same adult is involved in several POPs— so there's no bias, only a loss of precision.) Fortunately, (half)sibs do not seem to be common in our juvenile samples, and for clarity we therefore ignore the possibility of (half)sibs in the discussions below.

There are other phenomena that might at first be suspected of causing overdispersion, but careful thought is required. For example, the 38,000,000 SBT comparisons are based on "only" 13,000 fish, each being used in multiple comparisons. Does this somehow mean that the "effective sample size" is much smaller, i.e. that there is somehow serious non-independence amongst the 38,000,000 comparisons? No— but the reasoning is subtle. Ignoring sibs as per above, consider a comparison of two fish, juvenile J and adult A, in the "cartoon" version. With no further information except the population size N, the chance of a POP would be 2/N. Assume (as with SBT) that N is large, the sample is moderately large, and the number of POPs is small. Independence amounts to the following question: does knowing that (i) J is not in a POP with any of the *other* non-A adults, and (ii) A is not in a POP with any of the *other* non-J juveniles, help us to predict the outcome of the J-A comparison?

The information in (ii) is irrelevant (given that the other juveniles aren't halfsibs of J), because if N is large then the number of non-J offspring of any adult in the sample will almost always be zero anyway, so knowing that it really is zero for one particular adult is not informative. And as for (i): knowing that the other sampled adults aren't J's parents tells us almost nothing almost nothing about whether A will be J's parent[23]. Finally, comparisons that don't involve either J or A are obviously irrelevant. So, at least in the more than 98% of comparisons that don't involve a member of a true POP, knowing the result of all the other comparisons doesn't help us predict the outcome of this one— which is the definition of independence. [If the sampled fraction of fish was a substantial proportion of the total population size, and/or if a substantial proportion of the sampled fish turned up in POPs, and/or if there were many sibs in the samples, this argument would break down.]

Another phenomenon that might superficially seem like a source of overdispersion but actually isn't, is the non-random sampling of juveniles, e.g. shifts in sampling locations within the GAB between years. Non-random juvenile sampling has in fact been a deliberate aspect of the design all along, from the 2007 CCSBT paper onwards; for example, we don't sample any juveniles off South Africa. However, as noted in that paper, the only things that matter in order to keep the comparisons statistically independent, are that (i) there are few (half)sibs among the juvenile samples, and (ii) that the *adults* be sampled randomly (apart from selectivity and other effects that are specifically allowed for in the mini-assessment). Even then, all that "randomly" has to mean is: "a parent of one of the sampled juveniles is just as likely to be sampled X years after that juvenile's birth, as is another adult of the same sex, age, and size".

There is one other phenomenon which theoretically could be important for CK abundance

[23] "Almost" because this information does slightly reduce the potential pool of parents, from N to [N minus the adult sample size].

estimates, not so much for overdispersion as for bias: an unholy trinity of cryptic stock structure, biassed sampling of adults, and biassed sampling of juveniles. A lengthy explanation was given in our 2007 CCSBT paper, and is copied below. The key point to add in 2012, is that we have now checked as suggested in 2007 for any *temporal* substructure on the spawning grounds (see 3.3), and found none; we have not checked *spatial* substructure, but as below this seems *a priori* unlikely.

[4.7 from CCSBT 2007 CK paper] Population structure

So far, it has been assumed that SBT form a single population with complete interbreeding. Although no previous study has found evidence of population structure, conventional population genetics applied to large populations is a notoriously blunt tool for that task. It turns out (see **[6.0.6]**) that the basic method is unbiassed even when there is population sub-structure, providing that sampling is proportional to abundance across either the sub-populations of adults, or the sub-populations of juveniles. In our SBT project, juvenile samples come only from the GAB, so if there are substantial numbers of non-GAB juveniles out there somewhere, then juvenile sampling will obviously not be proportional. However, adult samples should cover the spawning season and spawning area, although not necessarily in strict proportion to adult SBT density. Hence, the basic estimator would exhibit population-structure bias if and only if three conditions all apply:

1. adults exhibit fidelity across years to particular parts of the spawning season and/or spwaning grounds;

2. the timing or location of spawning affects a juvenile's chances of going to the GAB (rather than going elsewhere or dying young);

3. sampling coverage of the spawning grounds (in time and space) is substantially uneven, and correlated with the fidelity patterns in (1). (In other words, if adults showed timing-fidelity but not spatial-fidelity, whereas coverage was even across the spawning season but not across the spawning grounds, then the uneven spatial coverage would not matter.)

There is no direct information on condition 1. With respect to condition 2, much the greatest part of SBT spawning occurs within the North Australian Basin ([**?**]), and particularly towards the east and south of the basin beyond the Australian shelf, where the Indonesian through-flows in summer would tend to push the larvae together into the Leeuwin current. These conditions seem unlikely to induce a strong location-of-spawning effect on most juvenile's subsequent propensity to go to the GAB[24], although a timing-of-spawning effect is possible. With respect to condition 3, the Benoa-based operations that we are sampling coincide well with this main spawning area ([**?**], Figure 4.3.1; note that the fishing range has expanded southwards since then, as per [**?**]). Approximate timing-of-effort information could be probably be obtained from the sampling program; spatial information has proved harder to get, but the data obviously do exist somewhere at the company level, and some insights may be obtainable through, for example, the observer program ([**?**]) or the Fishery High School program ([**?**]).

Fortunately, there is enough information in the project data to check the first two conditions. If the seasonal/spatial distribution of identified parents of GAB juveniles is substantially different to the seasonal/spatial distribution of all adult samples, then that is a clear signal that

[24] A small proportion of larvae are found to the north of the NAB and west of it. Different oceanographic conditions apply there, and those larvae could well end up somewhere different as juveniles. However, at least until 1981, this proportion was small.

the first two conditions do apply. Such evidence of population structure[25] would be of major qualitative importance to management, regardless of its impact on quantitative results.

If and only if the first two conditions do apply, then the third could be checked using timing (and perhaps location) information on Indonesian samples. And if all three conditions do apply, then it should be possible to adjust for the uneven adult sampling probabilities, again using sampling coverage information. That is very much a bridge to be crossed only if we come to it; but because the sampling coverage is at least fairly complete[26] even if not necessarily balanced, we would in principle be able to develop a correction if required.

[6.0.6 from CCSBT 2007 CK paper] Population substructure and sampling bias Suppose the entire adult population of N is made up of two sub-populations with proportions π and $1-\pi$, and that adults are sampled proportionally from their respective sub-population, so that the overall adult sample contains $m_A\pi$ fish from the first sub-population and $m_A(1-\pi)$ from the second. Juveniles, though, are not necessarily sampled in proportion to sub-population abundance; let m_{J1} and m_{J2} be the numbers sampled from each sub-population.

If the entire dataset is analysed without regard to sub-populations, then the expected number of POPs can be calculated by considering samples from each sub-population separately (since there will be no cross-POPs between juveniles from one sub-population and adults from the other):

$$\begin{aligned}\mathbb{E}[H] &= \frac{2m_{J1}(\pi m_A)}{\pi N} + \frac{2m_{J2}(1-\pi)m_A}{(1-\pi)N} \\ &= \frac{2m_{J1}m_A}{N} + \frac{2m_{J2}m_A}{N} \\ &= \frac{2m_J m_A}{N}\end{aligned}$$

just as in the case without sub-populations. In other words, the basic estimate is unbiassed provided at least one life-stage is sampled in proportion to sub-population abundance. If both are sampled disproportionately, though, there will be bias.

[25] "Population structure" is probably the wrong phrase, because the behaviour does not have to be heritable; adult spawning preference need not be related to earlier juvenile GABness, even if offspring's GABness is driven by adult spawning preference.

[26] Again: over the great majority of the spawning area.