OCEANS AND ATMOSPHERE
www.csiro.au

SBT Close-Kin Mark-Recapture: options for the medium term

Contract report for CCSBT, 2015

Mark Bravington, Paige Eveson, Peter Grewe, Campbell Davies:

CSIRO, Hobart, Australia

SBT CLOSE-KIN MARK-RECAPTURE: OPTIONS FOR THE MEDIUM TERM CONTRACT REPORT FOR CCSBT

MARK BRAVINGTON, PAIGE EVESON, PETER GREWE, CAMPBELL DAVIES

JULY 2016

Abstract. Direct monitoring of adult SBT stock via Close-Kin Mark-Recapture can be conducted effectively (i.e. to give reasonably precise estimates of recent trends) by genotyping 1000 adults and 1000 juveniles per year from sampling in 2016 thru 2019. The back-catalog of samples from 2011-2015 also needs to be processed, which can be done by 2017 or by 2019 depending on urgency. Sample size requirements should be revisited in 2019, to reflect new information from the CKMR itself and the ongoing needs of CCSBT. To achieve useful CVs with these low sample sizes, the genotyping method needs be changed away from microsatellites to a more modern sequencer-based method capable of revealing Half-Sibling Pairs (HSPs) as well as the Parent-Offspring Pairs (POPs) which form the basis of our previous studies; having HSPs as well as POPs leads to more robust CKMR modelling and lower sample size requirements. We propose one specific new genotype method, a variant of that suggested last year. We have now thoroughly tested it on SBT, and the new method is very cheap (lower unit cost than any competitor), highly reliable for POPs, and fully capable of finding HSPs. Adopting the new method does mean that some existing samples from 2006-2010 will need to be re-genotyped, but the reduced sample sizes and unit costs will lead to overall cost savings by the time the 2011-2015 back-catalog has been genotyped.

Thanks to Eric Anderson and Robin Waples for review and discussion.

Contents

1. Introduction

Close-Kin Mark Recapture (CKMR) is a new approach to estimating adult abundance and demographic parameters. It uses modern genetics to identify close relatives (parents and offspring, or half-siblings that share one parent) amongst large sample sizes of fish, and then makes demographic inferences about the adult stock from the number and pattern of pairs found. The estimates are "fishery-independent", in that they do not need CPUE or total catch data, though they may require length and age compositions and/or supporting biological data. Thus CKMR can be used as an independent check on conventional stock assessments when there are concerns about the validity or interpretation of catch or CPUE data. CKMR data can also be incorporated into a stock assessment (e.g. the SBT OM) in much the same way as tagging data from individual mark-recapture studies.

The first application of CKMR has been to SBT, using genetic samples collected over 2006-2010 from adults in the spawning grounds off Indonesia, and 3yo juveniles in the Great Australian Bight. The method seems to have worked well (Bravington et al., 2014; CCSBT, 2012; CCSBT, 2013); it gave much better precision on the adult abundance estimate, ruled out certain scenarios previously considered by the OM that were inconsistent with the CK data, and showed that the age-specific pattern of reproductive contribution (ie the appropriate definition of "Spawning Stock Biomass") is quite different to what was previously assumed. The CK data have now also been incorporated into the SBT OM (Hillary et al., 2012).

CKMR clearly has potential not just as a one-off abundance estimation method, but as an ongoing data stream to monitor adult abundance. This is appealing because the SBT management target is to rebuild the spawning stock, and apart from CKMR there is no other proposed method of directly monitoring that stock. For example, a CKMR-based estimate of adult abundance would automatically reflect losses due to large IUU catches, something that is near-impossible to detect in conventional assessments which rely on known catches.

In 2014, CCSBT therefore funded a scoping study on using CKMR for monitoring SBT adult abundance. During that study, it became clear that changing to a newer genotyping method could provide a better basis for further SBT CKMR work: more informative, more robust, and cheaper. As well as delivering POPs (Parent-Offspring Pairs), some new genotyping methods have high enough resolution to also find HSPs (Half-Sibling Pairs) among juveniles, which offers considerable improvements over POP-only CKMR: more robust analysis, and lower sample size requirements (i.e., cheaper). Initial results were presented for one new genotyping method at the 2014 ESC, but there was not time to investigate fully. CCSBT therefore contracted a follow-up study for 2015, and this is the report.

The main differences from last years' study are:

- we have settled on recommending one particular new genotyping method— a variant of last year's, but even more accurate and substantially cheaper;
- we have thoroughly tested it on SBT, to show it can find not just POPs but HSPs too;
- the genetic work has been reviewed by external experts [to be provided as a separate document];
- we have modified the mini-assessment-model used for predicting CVs under various future sampling scenarios, making it more robust by relaxing some assumptions.

The report structure has also been simplified, with recommendations given first and the technical material deferred to Appendices: the final Appendix on genetics, in particular, is quite long.

2. Recommendations

2.1. **Genotyping method.** Microsatellites ("usats"), as used for genotyping in our initial CKMR study, are no longer the best genotyping method for CKMR. After considering numerous alternatives, we chose one for full evaluation this year, referred to here as "HFS-Dart"[1]; it is a refinement of the SNP-sequencing method called

[1]HFS = "Highly Focussed Sequencing". Dart = "Distributed Array Technologies", the company we have worked with to implement this genotyping method. See Appendix for details.

"ddRAD" in genetics literature. The results look excellent, and we recommend changing from microsatellites to HFS-Dart; costs will be more than recouped by the time the back-catalogue of existing ungenotyped samples (2011-2015) has been processed. While the nature of CKMR makes it fundamentally impossible to run a *small* pilot study to directly find HSPs[2], we have shown this year that HFS-Dart is easily accurate enough to find HSPs with SBT (Appendix C); in fact, we have already found HSPs in two shark species using a less-accurate version of the same method (the version discussed in our ESC 2014 report).

Microsatellites worked well for our initial study, so any proposed alternative method needs to pass a set of questions:

- Can it find POPs reliably?
  - Yes, unquestionably, for HFS-Dart. (Clearly demonstrated in our tests this year, using known POPs from the previous study.)
- Is it cheaper?
  - Yes. Unit cost of HFS-Dart is about half that of usats[3]. Changing technology would incur a one-off cost to re-genotype some existing samples from 2006-2010, but that cost would be recouped by savings in the 2011-2015 back-catalogue alone.
- Is it future-proof?
  - Yes. The ddRAD technology behind HFS-Dart is sequencer-based, and sequencing is at the core of modern genomics (unlike microsatellites); it is not going to go away. Although we have worked with one particular supplier, Dart Pty Ltd., and would strongly advocate staying with them, the same SNPs we have used could be genotyped by another ddRAD supplier.
- Does it do more than usats can?
  - Yes. With HFS-Dart it should be possible to find HSPs as well as POPs— not possible with usats or with other new methods such as targetted SNP assays. With HSPs we avoid the need for an untestable assumption about daily-catchability-at-size; also, the increased number of "CK-pairs-per-sample" means that the annual sample size (and cost) can be reduced relative to a POP-only approach (Appendix A).
    - Note that the unit cost of HFS-Dart would not be reduced further even if we just wanted to target POPs and therefore needed fewer SNPs. The preparation steps impose a fixed overhead, so the extra loci available for HSPs are in effect "free"; the recommendation to switch to HFS-Dart applies regardless of whether and how HSPs are actually used.
- Is anything even better likely to come along?
  - Genetics moves quickly; who knows what will be available in 5 years' time? However, CKMR overall is *already very cheap*, and the pure *genotyping* cost (with HFS-Dart) is not much more than 50% of the full cost of dealing with a sample, and there are fixed costs for statistical analysis etc., so the opportunities for future savings are restricted regardless of technology. POPs especially with HSPs are entirely adequate for CKMR monitoring, so, provided that a technology passes the tests above— which we consider that HFS-Dart does— there is no point to waiting for a hypothetical "even better technology": HFS-Dart is good enough and cheap enough to adopt now.

2.2. **Sample sizes and strategy.** Based on the detailed simulation results in section 3, we recommend an annual sample size for genotyping of 1000 juveniles and 1000 adults over the next few years. This would allow CVs of 25% on 10-year trend estimates[4] by 2017, dropping to 20% by 2019; CVs on point estimates would range from 20-30%

[2]For example: 10% of the total sample will only contain 1% of the total HSPs, not enough for a reliable test. This is the flipside of the "quadratic magic" of CKMR, whereby the more samples you collect, the higher the *proportion* that are involved in CK pairs. Long-term large-scale CKMR studies are relatively cheap, but small-scale "rough" studies are relatively expensive.

[3]That is purely for genotyping; the other steps in collection/processing etc. cost the same regardless of method.

[4]Not strictly a CV: see notes on Table 1

depending on how recent the year-of-interest (the more retrospective, the better the CV). The results assume that HFS-Dart genotyping is adopted, so that HSPs (or at least some proportion of them) can be identified; CVs would be considerably higher otherwise, or sample sizes would need to be increased.

Longer-term sampling strategy should be revisited in 2018 or 2019, when enough results are available; in particular, the story told by the HSP data will affect the best options for the future, and may allow a "tighter" model (less conservative assumptions) with correspondingly lower CVs.

Note that ongoing collection of Indonesian otolith sampling is essential for achieving these CVs, and will become more so in future as the CKMR assessment evolves to allow changing selectivity.

The genetic sampling falls into three categories:

Re-genotyping:: Sample-pairs cannot be checked for POP status unless the same genotyping method is used for both. Therefore, a change of genotyping method may entail re-genotyping some or all of the 2006-2010 samples, which have already been genotyped with usats. Re-genotyping is fairly cheap, since the preliminary DNA-processing steps are already complete. Adult and juvenile samples need to be considered separately; adults have limited "shelf-life" since it does not produce offspring after it is caught, whereas a juvenile continues to "tag" its parents indefinitely even after its death. The simulations suggest that it **is** worth re-genotyping juveniles from 2006-2010, but **not** adults from those years.

Back-catalog: of samples from 2011-2015 also need to be genotyped; we have used the same annual sample size of 1000 adults and 1000 juveniles. The back-catalog is slightly cheaper to genotype since no collection costs are entailed. This could be completed by the 2017 ESC.

Future samples: from 2016 onwards, at 1000 adults and 1000 juveniles per year. There is no clear value at present to vary the ratio of adults to juveniles; this should be reconsidered once results from the next few years are available. It is desirable to *collect* more than 1000 adults and juveniles per year (at least at the current level of 1500 of each), though, for two reasons:

- Collecting samples for CKMR is much cheaper than genotyping them. Having the samples "in the bank" leaves the option open in future to retrospectively increase the sample size, by genotyping archived samples.
- Subsampling may become desirable to concentrate on the fish of most interest. Specifically, this would give us the flexibility to:
  - genotype only juveniles that are sure to be 3yo (juvenile length-at-age distributions do shift from year to year, and knowing juvenile age is important for CKMR).
  - deal more easily with changing selectivity in the Indonesian fishery[5]. Fish younger than 10yo have lower individual annual fecundity and so less per capita chance of being in a POP; it may be more efficient to concentrate on genotyping larger fish, as we did in the previous CKMR study.

3. Sample size and design

3.1. **Model framework.** The population dynamics model behind these results is the same as last year's "Approximate Mini-Assessment" (Appendix B): an age-based approximation to a length-based model, assuming constant (but unknown) selectivity over time[6], no trend in 8yo recruitment, mortality fixed for plus-group at 0.29 and estimated for younger adults. CVs are computed using simulated CKMR results (POPs and HSPs from pairwise comparisons) combined with a "age-based summary" of Indonesian catch compositions. No pre-adult data are used.

[5]Note that it is not essential for CKMR whether "adults" are actually caught on the spawning ground or elsewhere: CKMR only considers their reproductive contribution in years *before* they were caught, and the key point is rather that sampling of "adults" should be random except for selectivity. However, selectivity in the Indonesian catches is linked to the ability to estimate adult mortality, which is important for CKMR itself and of course for projections of future spawning stock.

[6]In our previous CKMR study, selectivity is assumed to be constant on average over time, with variations from year-to-year modelled as overdispersion in the length data. While that is a reasonable approximation in a short-term study, it is not necessarily robust over longer periods. This could be resolved in a real analysis, especially with HSPs to help separate selectivity/mortality/annual fecundity, but is too complicated for a design study.

It is a simplification of the CKMR standalone assessment in Bravington et al., 2014, expanded to optionally handle HSPs as well as POPs. When analysing the real data in the future, a more complicated length-and-age-structured model should/will be used, but for design purposes the simpler version shown here should be adequate. However, there are some caveats about the way the length/age data are "summarized", which seem to have the effect of making long-term CVs look somewhat good and short-term CVs somewhat bad, compared to the best available version of "truth"— the previous stand-alone CKMR mini-assessment.

In any case, the achieved CVs will inevitably be different because (i) current reality may not be exactly what the OM/CKMR leads us to think, (ii) future reality will surely drift away from "average projections", and (iii) the future data will themselves suggest ways to improve the model. Such things are inevitable in any design study, and the general corollary is not to cut corners on sampling, because CVs may turn out to be somewhat worse than the study suggests (of course, they might also turn out better, but that is not a problem). In the case of CKMR, a safeguard is available by collecting more samples than are initially genotyped.

3.2. **Results.** Table 1 shows expected CVs under various scenarios (i.e. different sample sizes and/or analysis options) and at various review dates. Each scenario is the same as the one above it, except where a new parameter value is shown; for example, in the 2017 table (top) scenario 3 is like scenario 2 except that the juvenile sample size is 1000 and the adult sample size is 500 per year. "Sample size" here refers to numbers genotyped, not numbers collected, which may and should be larger. All scenarios assume that the current 2011-2015 backlog of samples has been genotyped (at the same numbers-per-year shown in the Tables) by the review-year in question. For brevity, the only results shown are CVs of retrospective point estimates of "N10+", the number of adults aged 10 or above. Since trends may be of as much interest as absolutes, the SE of change-in-log-adult-abundance between 2006 and 2016 is also shown (multiplied by 100 and effectively on a "%" scale, comparable with CVs). For reference, a true doubling/halving of abundance over that period would correspond to a true change of about +/-70, so would be easily distinguishable from noise if the SE is 20 "%"; however, presumably CCSBT is also interested in changes that are considerably less drastic.

For reference: the only differences from the simulations in last year's report are:

- a 1-year shift of review dates (to coincide better with CCSBT schedules)
- the default assumptions about HSPs are more conservative this year. Specifically:
  - the "separate N for HSP" trick is used to guard against "non-GAB breeders" and any other unmodelled heterogeneity in reproductive output (section A.2).
  - not all HSPs may be identified, and a predictable proportion may be lost below the threshold (section C.12.2)

These more conservative assumptions of course give higher CVs for a given sample size, so the proposed sample sizes are higher than last year.

TABLE 1. Retrospective CVs (%) at review-years 2017, 2019, 2025 (top to bottom), under different sampling and modelling scenarios. "SE_trend" refers to estimated change in log(N10+) between 2006 and 2016.

	PARAMETERS
Mju (Mad)	Annual number of genotyped juveniles (adults) from 2011 onwards
HSP	¿Use HSPs in the modelling?
pHSP	Proportion of HSPs assumed reliably identifiable
Tredoju/ad	Regenotype (HFS-Dart) juveniles/adults samples starting from this year
link	(If no HSP) ¿Assume selectivity and fecundity are linked via "constant-daily-catchability"?
separate	¿Treat "N for HSP" separately to "N for POP"?
juage	Safety option if uncertain about juvenile age in Port Lincoln samples

2017 review

	Mju	Mad	HSP	Tredoju	Tredoad	pHSP	link	separate	juage	CV_2010	CV_2017	SE_trend	E[HSP]	E[POP]
1	1000	1000	Y	2006	2011	0.75	N	Y	N	20	31	24	190	111
2	500									23	38	32	111	91
3	1000	500								24	36	26	190	78
4					2009					23	36	26	190	84
5				2011	2011					30	53	41	52	65
6	1500	1500	N	2006				N		18	30	28	0	175
7							Y			11	26	28	0	175
8			Y				N	Y		16	25	19	296	175
9								N		10	22	19	296	175
10						1		Y		16	24	17	395	175
11									Y	17	26	20	298	175

2019 review

	Mju	Mad	HSP	Tredoju	Tredoad	pHSP	separate	juage	CV_2010	CV_2017	SE_trend	E[HSP]	E[POP]
1	1000	1000	Y	2006	2011	0.75	Y	N	18	26	18	233	132
2	1500	1500							14	20	14	383	219
3						1			14	19	13	511	219
4								Y	15	21	15	396	219

2025 review

	Mju	Mad	HSP	Tredoju	Tredoad	pHSP	separate	CV_2010	CV_2017	CV_2024	SE_trend	E[HSP]	E[POP]
1	1000	1000	Y	2006	2011	0.75	Y	15	18	16	10	349	193
2	1500	1500						11	13	12	8	635	355
3						1		11	13	11	7	847	355
4							N	6	8	9	7	847	355

A number of conclusions can be made, as follows. Line (scenario) number implies a comparison with the scenario immediately above it in the topmost (2017-review) Table, unless otherwise stated; these comparisons are in terms of "large" effects on CV.

- There is little point in re-genotyping 2006-2010 adult samples with HFS-Dart (line 4), which would only close a small gap (for comparisons with juveniles caught between 2011 and 2013— after 2013, only post-2010 adults are informative anyway); note that there is no point in re-genotyping adults caught prior to 2008 since any offspring would have already been genotyped.
- However, it is worth re-genotyping 2006-2011 juveniles (line 5 vs 3)
- Without HSPs, the residence-time-based link between selectivity and fecundity link is a powerful assumption that *if correct* leads to substantial lower CVs (line 7)— but it cannot be tested without HSPs.
  - With HSPs, that assumption can be relaxed without hurting the CV (line 8 vs 6)
  - If there turns out to be no reason to fit separate N for HSPs to POPs (section A) then the CV can be improved very considerably (line 9; also line 4 of 2025 Table)
- The estimates can be made robust to (some) uncertainty in juvenile age— there is little impact on CV of taking the safe/robust approach of avoiding comparisons between juveniles caught 0 or 1 years apart, to make sure that there are no within-cohort comparisons (line 11; also line 4 of 2019 Table).

- Results are not highly dependent on "pHSP", the proportion of HSPs that can be reliably identified (line 10 vs 8; also lines 3 of the 2019 and 2025 Tables). This is at first surprising, because HSPs overall are important and a 25% change in the number found represents a lot of information. The explanation is probably that HSPs now get their "own" N, so the number of HSPs found primarily inform the "HSP N" rather than the overall N. However, the pattern of HSPs found over time certainly informs other parameters such as fecundity, so the extra HSPs available if pHSP turns out to be higher may provide extra information on spawning stock trends beyond N-trends.
- Slightly-retrospective estimates are considerably more precise than fully-up-to-date estimates. Compare the CV_2017 columns in the 2017 and 2019 Tables; the latter are substantially more precise, more so than would be expected just because of two years' additional samples by 2019. The effect on 2006-2016 trend precision is similar.
- There is quite a substantial difference in CV between total 2000 and 3000 genotypes per year, at least up to 2019 (line 8 vs line 1 of 2017 Table; line 2 vs line 1 of 2019 Table).
  - Reducing the number of juveniles is worse for CV than reducing the number of adults, in terms of ability to assess trend (lines 2 and 3); neither is a good idea.
- Overall, HSPs are very useful. In the "worst case", they can be used to relax otherwise-untestable assumptions (e.g. line 8 vs 6) without impacting the CV. And if the various assumptions (some of which, e.g. constant-daily-catchability, are currently embedded in the OM) do turn out to be valid, then HSPs would allow a given target CV to be achieved at substantially lower sample sizes (implicitly from line 9 vs 6; sample sizes are the same, but the line 9 CV with HSPs is much lower than the line 6 CV without).
- An equal mix of juveniles and adults is reasonably efficient; the optimal mix does depend on which model is appropriate (e.g. on whether a separate "HSP N" is needed) so at this stage it seems premature to consider varying the ratio away from 1:1.
- Several of the questions about appropriate models/parametrization ("separate", "pHSP", "juage", and even "link") will presumably be resolved after a few years' data are analyzed, when results under different models can be compared. Appropriate sample sizes for the longer term (say, beyond 2019) cannot really be determined yet, for several reasons:
  - the constant-selectivity assumption used in this medium-term model becomes less and less realistic over longer time periods;
  - precisely which set of assumptions turns out to be appropriate, will have a major bearing on what sample size is required long-term;
  - the long-term way in which CKMR results get used in ongoing SBT management has yet to be considered.

The "base case" (lines 1 of the 2017 and 2019 tables), with 1000 adults and 1000 juveniles per year, using HSPs and with regenotyping juveniles only from 2006-2010, seems likely to provide useful CVs in a reasonably efficient way; we have used this as the basis for the indicative budget in the next section. However, analysis along the way may indicate that higher sample sizes are needed. Provided that we continue to *collect* more samples annually than we expect to genotype, then there is a buffer of "spare" samples available to allow "retrospective increases" in sample size.

4. Indicative budget

Table 2 shows likely costs for CKMR thru to 2019, under the "base case" proposal of the previous section.

- The ongoing annual cost will be about $200K, post 2019 when methods development is complete and way-of-use in CCSBT is settled (assuming an ongoing sample size of 2000/yr).

  - About $50K of the $200K comes from Indonesian otolith collection and age-reading— essential not just for CKMR but surely for any assessment of SBT "stock status".
- "Methods and database development" is front-loaded to 2016-2017, since we need to get a reliable storage and analysis system in place early on. It covers:
  - finalizing the genetic protocols;
  - formalizing the QC checks on genetic results, and the process for finding POPs and HSPs (based on work done this year in Appendix C, and leading to something resembling our existing "pipeline" for microsatellite genotyping);
  - systematic and transparent database processes for archiving the samples, genotypes, CK pair-finding, and abundance-estimation steps;
  - extending the length-and-age based "mini-assessment" (Bravington et al., 2014) to accommodate at least: HSPs as well as POPs; changing selectivity in the Indonesian fishery; possibility of length-based-ageing error in juveniles.
  - micro-management of subsampling for genotyping, etc., and consideration of long-term sample sizes in the light of CKMR results from 2016–2019.
- The Table shows 2011-2015 back-catalog processed in time for ESC 2017, so that a 10-year trend estimate can be made that year. The same total cost for the back-catalog ($342K) could instead be spread over 4 years (so that estimates are available at ESC 2019), but obviously that would limit what could be said by ESC 2017.
- Unit costs are lower for re-genotyping 2006-2010, because all the preliminary steps of physical preparation are already done; intermediate for the back-catalog of 2011-2015, because the samples are already collected but still need physical preparation; and higher for future (2016+) samples, which also need to be collected. Some savings on the preparation steps may become possible after 2016. Genotyping costs include $5K per year to process mtDNA from identified HSPs, so that maternal/paternal descent can be established (section A.4).

TABLE 2. Indicative budget; see text for notes. All costs in $1000 AUD

Component	Activity	2016	2017	2018	2019
Methods and database development	(see text)	142	146	67	34
Re-geno 2006-2010	7448 juves: Genotyping only	72	72		
Geno back-catalog 2011-2015	Subsampling DNA extraction Genotyping (10000 samples)	171	171		
Ongoing collection of tissue and juvenile otoliths	Collection (2000 adults, 1600 juves/yr) 100 juve otoliths	47	48	49	50
Geno ongoing	(Subsampling) DNA extraction Genotyping (1000 adults, 1000 juves/yr)	109	95	96	97
Indonesian otoliths	Sample & archive 1500 otoliths/yr; age 500	47	48	49	50
TOTAL		588	580	261	231

Appendix A. Uses of HSPs in CKMR

This material is taken from last year's report, with one significant change. It is possible that some adults give birth to offspring which *systematically* tend not to go the GAB in summer as juveniles. Since all the juvenile samples come from the GAB, such adults would be invisible in an HSP-based estimate of adult abundance, which would thus be negatively biassed, regardless of whether that propensity is heritable. In fact, any source of systematic inter-individual variation in reproductive output that is *not* described by length, age, or survival (which are explicitly accounted for in the model) would act as a source of negative bias in the HSP-based estimate— though, given the lifestyle of SBT, it is not obvious where such unmodelled heterogeneity might come from. In any case, any potential problem can be alleviated by estimating a separate N-for-HSP which is proportional (by sex) to the corresponding true adult N at age and year.

A.1. **Why HSPs?** Setting aside cost and logistic issues for now, there are two main reasons why HSPs as well as POPs are desirable:

- HSPs provide more CKPs per sample, thus addressing one of the main limitations on CV of abundance estimates.

In fact, HSPs will probably triple the total number of CKPs per sample, allowing us to substantially cut the long-term sample size requirement for monitoring.

- With HSPs as well as POPs, we become able to separately estimate (adult) selectivity-at-size, fecundity-at-size, and mortality.

Until now, with POPs alone, we have had to *assume* a fixed relationship between selectivity-at-length and fecundity-at-length, as determined by external data on daily fecundity and assuming that daily catchability is independent of size. This assumption has substantial implications e.g. for assessing trend in "effective SSB", so it is highly desirable to be able to avoid relying on it. (And of course the assumption is fundamentally incompatible with changes in adult selectivity over time.)

A third, and less important, benefit of HSPs, is that the presence of *large* numbers of *within-year* full-sibs[7] and half-sibs would[8] lead to overdispersion in the POP-based CK data. The evidence to date suggests not (e.g. so far no adult has been involved in more than one POP), but the only direct way to check is to actually look for half-sibs.

The rest of this subsection gives a informal perspective on the utility and potential pitfalls of using HSPs. The rest of section A takes a more formal approach, and also addresses the technical viability of finding HSPs; most of the material is deferred to Appendices.

A.1.1. *The cartoon version of CKMR.* The cartoon version (i.e. heavily simplified, mainly by ignoring time altogether) of CK MR with POPs goes like this. We start by genotyping one specific Juvenile. It must have two parents. If a random Adult is picked from the adult population of size N, then the chance of it being one of the parents of the Juvenile is $2/N$. If we repeat this exercise for each pairwise comparison between m_J juveniles and m_A adults, then the *expected* number of POPs found is $2m_Jm_A/N$, so N can be estimated from $2m_Jm_A/P$ where P is the *actual* number of POPs found.

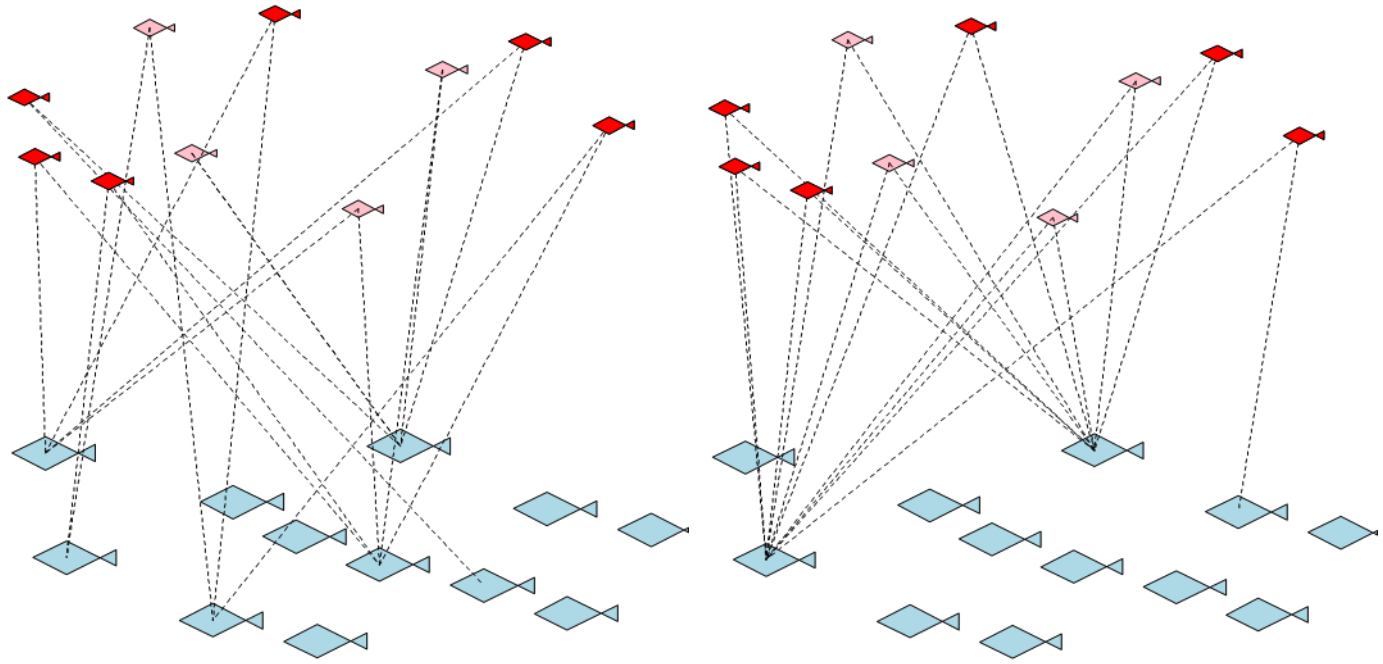
The analogous cartoon version with HSPs goes like this. We start by genotyping two Juveniles— John and Jane. They each have one mother. The chance that they have the *same* mother is 1 over the number of *female* adults, i.e. $2/N$. Similarly, the chance that they have the same father is $2/N$, so overall the chance that John and Jane are an HSP is $4/N$. Now, if we compare all m_J samples with each other, giving $m_J^2/2$ "non-double-counted" comparisons, and find H HSPs in the process, then we can estimate N from $2m_J^2/N$.

[7] Identifying full-sibs is much easier genetically than identifying half-sibs, so if we can do the latter then we can certainly do the former.
[8] It is inevitable that there will be *small* numbers of HSPs within each cohort of juvenile samples, but only of the same order of magnitude as the proportion in POPs, which is very low.

When sampling is split 1:1 between juveniles and adults[9], the cartoon suggests that the number of POPs and HSPs should be about equal, but for many reasons this need not apply in practice. Most notably, after an adult has been caught, it can no longer produce offspring, so the POP data-series is "one-way" in time; in contrast, catching a Juvenile does not stop its parent from producing more half-siblings of that Juvenile. The simulations for SBT shown later in this document indeed suggest that about twice as many HSPs as POPs would be found under 1:1 sampling.

A.1.2. *Variation in fecundity/reproductive success.* A FAQ about POP-based CKMR is: but is it affected by variation in reproductive success? The short answer, at least in the cartoon world, is "no" (Figure A.1); there is no bias, though variance would be affected by high RV since it would no longer be true that the comparisons are (almost) statistically independent. Having said that, in the non-cartoon world of SBT, fecundity (which varies systematically through an adult's life, as it grows) *is* important in analysing CKMR data, and can be allowed for; but year-to-year fluctuations in reproductive output make no difference.

FIGURE A.1. Non-impact of reproductive variability on POP-based CK MR



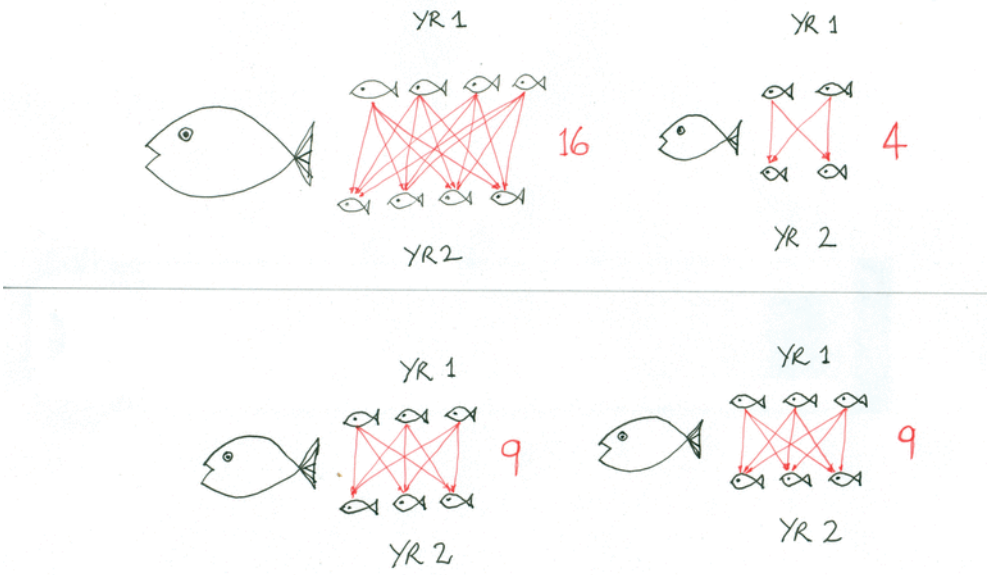
Low RV on the left; high RV on the right. However, the number of links in both pictures is the same. Each pairwise comparison has an equal chance of finding a link, and the estimate of adult abundance is based on the number of *links* found, not the number of *parents* found.

With HSPs, though, variations in reproductive output is much more important. Figure A.2 shows how systematic variation in fecundity (i.e. *average* reproductive output per year) affects the number of HSPs present. There are 20 cross-cohort HSPs (ie only comparing juveniles in *different* cohorts) in the top scenario, but only 18 in the lower. The same is true within-year; in the Figure, there are 30 comparisons per row (within-cohort) of juveniles, yielding 21 within-cohort HSPs in the upper scenario, but only 16 in the lower.

It is certainly possible that random within-year fluctuations in individual reproductive success could affect the number of HSPs found within each sampled cohort of juvenile SBT, so to be on the safe side we exclude within-cohort comparisons when using HSPs in formal CKMR.

[9]In this paper, a ratio M:N always means "future ratio of genotyped juveniles to adults"]

FIGURE A.2. Impact of reproductive variability on HSP-based CKMR



Variable fecundity scenario on top; constant fecundity below. Bigger fish make more babies. The upper and lower sets of juveniles *within* each scenario represent different cohorts.

Further, if there is some concern that age might not be accurately estimated in juveniles (note that the SBT juvenile samples are assigned to age-3 based only on length, not otolith readings), then an extra level of safety can be obtained by not comparing juveniles caught in year y with other year-y juveniles, nor to year-$y+1$ juveniles (assuming that any ageing error is one-sided, which for juvenile SBT would be the case; in some years there might be slight overlap between "typical 3yo size" and either 2yo size or 4yo size, but not both). There might still be one-year ageing errors within a few HSPs, but this is not so serious; however, a substantial inflation of the *number* of HSPs due to within-cohort reproductive variability (even from a small number of accidental within-cohort comparisons, if the variability is large) would be much more serious.

Even in the cartoon world, that still leaves growth-related fecundity to be dealt with. There is no way to allow for fecundity using HSPs on their own (because HSPs give no information on which adults were responsible for them), but of course for SBT we have the POPs as well, from which the age- (or size-)specific fecundity can be estimated. Roughly speaking, this allows "independent" estimates of abundance to be made from the number of HSPs, as well as from the number of POPs. More accurately, the two abundance estimates are only independent once the structural parameters (fecundity, mortality, selectivity) are estimated, as described in the next subsection. The combination of the two types of CKP is more powerful than either alone.

A.1.3. *A likelihood framework for HSPs.* With POPs, the non-cartoon version of estimation is as follows. The likelihood (a function of the model parameters) is formed from a set of pairwise comparisons, each treated as an independent Bernoulli trial (outcome yes or no). For adult i and juvenile j that are about to be compared, we first calculate the probability that i is j's Mother (assuming we know that i is female). If K_{ij} denotes the kinship-type of i and j (PO or UP— which here includes all other types of kinship), z_i the measured data on i, y_j the year-of-birth of j, and RO_{iy} the Reproductive Output of i in year y (measured in terms of juveniles that subsequently survive to j's age at capture)

$$p_{ij}^{\mathrm{PO}} = \mathbb{P}\left[K_{ij} = \mathrm{PO}\right] = \frac{\mathbb{E}\left[\mathrm{RO}_{iy_j}|z_i\right]}{\mathbb{E}\left[\mathrm{RO}_{♀y_j}\right]} \tag{A.1}$$

where the denominator is summed over all females alive at y_j (according to the model). The numerator involves y's expected size at y_j (since it was clearly caught after y_j). Fecundity-at-size parameters affect both the numerator and the denominator.

Having computed p_{ij}^{PO}, we compare genotypes, and compute the log-likelihood for that comparison as with any Bernoulli trial:

$$\Lambda_{ij}^{\mathrm{PO}} = \mathbb{I}\left[K_{ij} = \mathrm{PO}\right] \log\left(\frac{p_{ij}^{\mathrm{PO}}}{1 - p_{ij}^{\mathrm{PO}}}\right) + \log\left(1 - p_{ij}^{\mathrm{PO}}\right) \tag{A.2}$$

The overall log-likelihood is formed by summing over pairwise comparisons, as well as adding extra terms e.g. for the observed length- and age-distributions of sampled adults; see Bravington et al., 2014 for more details. Although comparisons are not strictly independent, treating the pairwise sum as an approximate likelihood is valid (to first order) provided that sampling is "sparse" (i.e. that the sample size is a small fraction of abundance that year— certainly true for SBT) and that sibship within juvenile samples from the same cohort is not too high. Detailed analysis of the usat results in fact shows that such sibship cannot be very high (say, no more than 10% of samples), otherwise the fit between observed and expected "near-misses" of POPs would not be as close as it is.

The case of HSPs is quite similar. For simplicity, consider just maternal HSPs. If the comparison is between two juveniles j and r— by assumption, born in different years, so take j to be younger than r— then we need to sum over all potential mothers of j:

$$\mathbb{P}\left[K_{jr} = \mathrm{HS}\right] = \sum_{\mathcal{F}_{y_j}} \mathbb{P}\left[j\text{'s mother was } i\right] \times \mathbb{P}\left[r\text{'s mother was } i\right] \tag{A.3}$$

where $\mathcal{F}_{y_j}$ is the set of females alive in y_j (according to the model). Each such female has relevant covariates z_{iy_j} which are specific to that year, and known to the model; z_{iy_j} mostly describes annual fecundity, but also implicitly includes the fact that i is alive at y_j. The first term is

$$\mathbb{P}\left[j\text{'s mother was } i\right] = \frac{\mathbb{E}\left[\mathrm{RO}_{iy_j}|z_i\right]}{\mathbb{E}\left[\mathrm{RO}_{♀y_j}\right]} \tag{A.4}$$

and the second term is

$$\mathbb{P}\left[r\text{'s mother was } i\right] = \frac{\sum_z \mathbb{E}\left[\mathrm{RO}_{iy_r}|Z_{iy_r} = z\right] \times \mathbb{P}\left[Z_{iy_r=z}|z_{iy_j}\right]}{\mathbb{E}\left[\mathrm{RO}_{♀y_r}\right]} \tag{A.5}$$

where the numerator reflects that i's fecundity may have changed between y_j to y_r, and that i may have died; the calculation therefore depends on survival, fecundity, and growth parameters, as well as on total abundance and age/size composition (in the denominators. The pairwise log-likelihood follows analogously to the POP case.

The only implicit assumption behind eqn (A.3), given that within-cohort comparisons are excluded, is that RO is independent between years *given* the variables that are in the model (in this case fecundity, growth, and survival); see the next section for comment. Thus the pairwise sum $\Lambda^{\mathrm{HS}} = \sum_{jr} \Lambda_{jr}^{\mathrm{HS}}$ does constitute an "unbiassed estimating equation" in the statistical sense, and as such is a valid basis for unbiassed estimates of abundance and other parameters. As with POPs, though, one further condition is still required for the pairwise sum Λ^{HS} to be a valid approximate log-likelihood (specifically, for its Hessian to be approximately the Fisher information). That is: sibship rates (half or full) *within samples within cohort* must be low, otherwise comparisons are not independent. Each j is compared not just with r but also with r^* from the same cohort as r, so if r and r^* happen to be half-sibs

themselves then the two comparisons with j are not independent, and the variance of Λ^{HS} will not be equal to the sum of the variances of all the $\Lambda^{\mathrm{HS}}_{j\cdot}$.

Although this cannot be too serious for SBT (for the same reason as with POPs), with HFS-Dart we can actually check the within-cohort-sample sibhsip directly (we can of course do the within-cohort comparisons, but not include them in the likelihood). If the proportion is substantial, it is easy to calculate the variance of $\sum \Lambda^{\mathrm{HS}}_{jr}$ accounting for non-independence of a comparison from cohort A to samples from cohort B, and to use this to rescale Λ^{HS} to have the correct first-order properties of a true log-likelihood.

As pointed out by one reviewer, there is a close link between HSP-based approaches to estimating census population size, and effective population size (e.g. WANG, 2009), which deserves further exploration. N_e depends on the variation across individuals in actual reproductive output, which is affected by both systematic and "random" terms (e.g. actual variation in lifespan). In a CKMR setting with pairwise comparisons for HSP, the systematic variation *should* be removed by taking account of whatever systematic effects can be modelled, as in eqn (A.5), so that "the N" in question is definitely "census N". However, omitting important systematic effects will certainly bias the resulting estimate away from "census N" downwards towards (some measure of) N_e. The next section mentions one possible cause for SBT, and a solution.

A.2. **Persistent offspring destination.** Although stock structure in the classical sense is not a concern for SBT, there remains at least a theoretical possibility of "persistent breeding destination", e.g. if the offspring of "Type I" females generally go to the GAB, whereas offspring of "Type II" females generally avoid the GAB. Then an estimate of N_{adult} constructed only from GAB-caught juveniles (which all of our samples are) would only be estimating those females of "Type I", and would be a negatively biassed estimate of total female adult abundance. Note that this applies even if "Type" is not heritable (so that classic stock structure does not exist); it only needs to apply for (adult) life.

This can be detected by allowing a separate N_{adult} to be estimated for HSPs alone. If this comes out substantially lower than the N_{adult} estimated for POPs, which is the real quantity of interest, then it is evidence against current assumptions (although it does not necessarily follow that there were many *surviving* juveniles outside the GAB in summer). Equivalently, this can be thought of as a sex-specific but time-invariant "scaling parameter" between N-for-HSP and N-for-POP that reflects *any* unmodelled heterogeneity across individuals in average reproductive output across years. Fecundity, growth, and survival are already in the model, so it is not obvious what might lead to substantial remaining heterogeneity. Such heterogeneity has been documented in some bird species (successful rearing can be a skill that takes years to learn, if ever), though it is not obvious how this might work for a species like SBT where the breeding behaviour is simply to spawn eggs or sperm for several weeks into a pelagic environment. Nevertheless, the extra scaling parameter is a safeguard, and it could subsequently be relaxed if analysis shows that it is close to one.

A.3. **Independent estimation of selectivity, fecundity, and mortality via HSPs and POPs.** The interpretation of POPs (ie the number found, and their patterns with age and time) is affected not just by adult abundance, fecundity, and mortality, but also by the complication of adult selectivity. To get round this for SBT CKMR, it has until now been necessary to assume that selectivity (within sex) is directly proportional to residence time on the spawning grounds, ie that catchability-per-day is independent of length (within sex). By making this assumption, and combining it with external data on daily spawning output as a function of (female) size, we enforce a "hard link" between selectivity and fecundity which allows them both to be estimated.

While it is hard to argue with the notion that residence time on the spawning grounds must be a primary driver of selectivity for SBT caught on the spawning grounds, it is also hard to argue that it must be the *only* important driver. The problem is that there has been no data to help; it *might* eventually be possible to e-tag enough adult SBT (recaptured later through the Indonesian fishery) to determine residence time directly; this has recently been

done for Atlantic Bluefin Tuna by Aranda et al., 2013, although unfortunately they could not accurately record body size. However, for SBT such a tagging program would take a long time and be expensive.

The basic ingredients for estimating structural (non-abundance) parameters from SBT POPs are age compositions (in adults, and in identified parents), and the mean time-delay between catching an offspring and its parent. It turns out that adult selectivity and mortality are inextricably linked in all these. In contrast, selectivity obviously does not affect HSPs at all. We therefore have some grounds to hope that HSPs and POPs together might be able to disentangle the parameters, without requiring the equal-daily-catchability assumption. The next paragraph gives a "heuristic" explanation of why this is possible.

Consider what happens after we catch one juvenile. As time goes by afterwards, how does the probability change that any one comparison reveals (i) a Parent, and (ii) a Half-Sibling? In both cases, the probability diminishes at the mortality rate because the parent might die, and also diminishes at the the population rate-of-increase because there are more non-Parents and more non-Half-Siblings around. Assuming that the parent does survive and is still around to be captured and/or make babies:

- For POPs: the parent becomes more catchable each year because of increasing selectivity with age
- For HSPs: more half-sibs are available for capture each year, because the parent becomes more *fecund* over time.

These trends in probability determine the mean time gap for POPs and for HSPs. The difference between the mean gaps for the two types of CKP thus reveals the difference between age-specific selectivity and fecundity (at least in terms of the overall trend with age). Once these two are distinguished, it becomes possible to estimate all remaining structural parameters including adult mortality, based on age compositions and the trend over time in CKPs-per-year.

Last year's report contains a formal proof that this separation is possible in a simple and abstract setting.

A.4. **Maternal vs paternal HSPs.** The expected number and pattern of HSPs depends on whether the shared parent is the Mother or the Father. To determine this, we can analyse the mitochondrial DNA (mtDNA) of all HSP fish. MtDNA is always inherited from the mother, so if the two HS have *different* mtDNA, they must have different mothers— ie their shared parent must be the father. On the other hand, if they have the *same* mtDNA, then they are likely to share a mother. However, there is a possibility that they actually share a father and just by chance happened to have different mothers with the same mtDNA. This probability $\mathbb{P}$[shared mother|HSP with same mtDNA] can be calculated from mtDNA allele (AKA "haplotype") frequencies in the population. Once the probability is known, the HSP data can be analysed statistically using mixture-distribution ideas (i.e without requiring certainty about which parent was shared).

For this to work well, the mtDNA genome must have reasonably high diversity, so that we can reliably identify a good number of true *paternal* HSPs. Even if there is enough real diversity, we may not detect it unless the mtDNA genotyping process is sensitive enough[10]. In Grewe's (1997) examination of SBT mtDNA, about 70% of fish all had particular one allele, i.e. an inconveniently low diversity. However, because the goal was traditional population genetics (and because of the analytical methods available at the time), the measurement technique was deliberately chosen *not* to discriminate finely within alleles. Modern techniques can discriminate much more finely, and we have recently done this successfully for a species with a genuinely low-diversity mtDNA (and nuclear) genome; based on what we know of the SBT genome, we expect that its true mtDNA diversity is acceptably high. As long as the frequency of the commonest *measured* allele can be reduced to below, say, 0.4, there should not be any great loss of precision introduced by uncertainty about which parent is shared in some of the HSPs.

Genotyping the HSP mtDNA to establish maternal/paternal origin may cost up to $50 per fish, but since it is only necessary for fish in HSPs which will amount to a few hundred individuals, the additional cost is negligible.

[10]Most genotyping of mtDNA uses only a small part of the mtDNA genome, and so will not differentiate between alleles (haplotypes) that differ only elsewhere in the mtDNA genome.

Appendix B. A simplified framework for designing SBT CKMR

[From 2014 report] To do the calculations for this report, it was necessary to implement a new "CK assessment model". The original CK Mini-Assessment (CKMA; CCSBT 2012 and 2013), which is fully length- as well as age-structured, is too complicated and slow for design purposes, has some structural constraints such as being unable to deal with changes in selectivity, and also gave baseline estimates with a much stronger trend[11] than the CK-OM combination model suggested. Since the recent trend has a big impact on future projections, it seemed wiser to start the projections with recent estimates from the CK-OM which (presumably) are better since they use other information. Some of the structural features of the Mini-Assessment may also be inadequate for long-term work, e.g. constant selectivity over time. Unfortunately, the CK-OM itself is not suitable for CK projections either; the most obvious problem is that it lumps both sexes together, despite plentiful sex data, clear evidence of dimorphic growth, and every reason to expect different selectivity and fecundity patterns by sex!

Instead, for this paper I developed a new Approximate Mini-Assessment (AMA) for adult population dynamics only, using just CK data and adult age data. It is purely age-and-sex-based model, with constant mortality z over time and over ages 8-24, a plus-group at[12] 25 with its own z_+, fixed sex-ratio over time, and sex-specific selectivity and fecundity curves.

To seed the model, I used the selectivity-at-length and fecundity-at-length estimates from the CKMA, converted from length to age. However, I also used the fitted age-composition data from the CK-OM, adjusted slightly in a least-squaresy way to match the detailed assumptions of the AMA (e.g. a constant z over ages 8-24, estimated from the CK-OM results). These choices are somewhat internally inconsistent; the CK-OM z is generally higher than the CKMA z, and would imply different selectivity and fecundity curves. I also adjusted the overall abundance to match the existing total of 45 POPs by 2010, since the untweaked AMA suggested rather more POPs (63) whereas, for purposes of assessing future CVs, we *know* we are starting with 45 POPs.

There is a fast and straightforward way to evaluate the information content of purely CK data; each pairwise comparison makes an independent contribution to the overall likelihood, and is in effect a Poisson-distributed RV with very low expected value, and its expected Fisher information ($d^2\text{loglik}/d\text{params}^2$) is easily enough computed. We need only simulate exact population dynamics, evaluate the probability of each comparison yielding a POP (or an HSP, as appropriate), count the number of comparisons of each type under a given sampling design, add up all the expected Fisher information matrices, and invert it to get the covariance matrix of all the parameters.

The more difficult problem is what to do about the Length- and Age-composition (L&A) data from Indonesia, which is *essential* to estimating abundance and everything else with CKMR data. In the CKMA, the L&A data were very informative, and allowing for uncertainty in them added remarkably little to the CV of abundance. Transition to a purely age-based setting is one tricky aspect, since only about 1/3 of adults are directly aged (although all identified parents are aged). In the AMA, age is assumed to be accurately known for all sampled adults (not just parents), so that every POP-seeking comparison is conditioned on adult age; although this is not realistic, all "real" adults are lengthed and sexed, and it is length rather than age that is likely to be the primary driver of selectivity and fecundity, so the "accurate age" assumption is a reasonable surrogate in a non-length-based formulation like AMA.

However, it seems unwise to treat age compositions as truly exact, in that it might interact subtly with detailed assumptions, e.g. about the shape of selectivity curves as opposed to the general trend with age, so as to provide "spurious information" about certain parameters. The compromise I used was to assume that the L&A data is accurate enough to reconstruct all the numbers-at-age-and-time (relative to some absolute abundance) *given* any set of parameters for mortality and selectivity. Thus the L&A data are assumed to provide *zero* information in

[11] The CK mini-assessment had considerable uncertainty about the *trend* in abundance over 2003-2010, though less so about the mean.
[12] To match the original choice in the CKMA, so that no extrapolation of previous results was required. With hindsight, 30 would have been a better choice in the original CKMA.

themselves about mortality or selectivity. However, once those parameters (and the overall abundance) are set, the L&A data entirely determine the numbers-at-age-and-time, from which the probabilities of finding CKPs are computed. Operationally, this is done by using the true selectivity (in the simulation) and the true population-numbers-at-age to fix the catch-proportion-at-age in Year 1, and the catch-proportion-at-age-8 (the youngest adults) in each subsequent year. Then, given some trial selectivity parameters during estimation (and a trial initial abundance), the catch-proportion-at-age in Year 1 is used to determine the corresponding initial numbers-at-age. These population numbers-at-age are projected to year 2 using mortality, and the incoming 8yo are set based on the catch-proporation-at-age-8 in year 2, and the selectivities. This process is projected forwards to fill in the entire numbers-at-age matrix. This *notionally* is supposed to correspond to the L&A data providing enough information to "back-converge" all cohorts to their initial relative abundances, but not to overfit to details of changing age distributions in the catch from year-to-year that result from the interaction of selectivity and mortality.

In the context of a "simple" model suitable for sampling design, I have been unable to think of any better yet practical way to handle the information associated with L&A, but it has to be admitted that this approach has flaws: it is too optimistic in some ways and too pessimistic in others. It is optimistic in the sense that, for making inferences near the end of the projections, the relative abundance of recently-recruited cohorts (ie cohorts currently not much older than 8) will *not* be known accurately; retrospective accuracy is only a reasonable approximation for cohorts that have been sampled as adults for enough years to accumulate reasonable sample sizes. [Having said that, in a "real" assessment that used pre-adult data as well, there would be prior information on the relative strength of each cohort as it nears adulthood.] By the same token, though,the AMA makes no assumption at all about trends in recruitment[13], which gives perhaps an unreasonable amount of freedom. Also, the AMA treatment of L&A also ignores detailed information about changing length-distributions which (given reasonable assumptions about length-at-age) may convey some real information about selectivity. But, on yet another hand, it is also not reasonable to assume over the long term that selectivity is constant, so in practice much of the information in detailed length data would be used to patch up short-term changes in selectivity. It is, in a word, complicated.

From experimentation, the L&A treatment leads to increased CVs relative to the CKMA in the short term (eg using only the existing data up to 2010). Then again, the CVs in the CKMA were remarkably close to the theoretical minimum imposed by the number of POPs found; perhaps some hidden structural assumption (e.g. about recruitment trends, or constant selectivity over time) was artificially lowering the CVs in the CKMA, and perhaps the CKMA CVs should not be over-trusted as a benchmark. Anyway, in the long term (20+ years), CVs from the AMA actually become implausibly low, this time for a clear reason: given the parameters for selectivity, fecundity, and mortality, only one single abundance parameter is required, and thereafter the L&A data is assumed to be capable of accurately setting the *entire* time-series of numbers-at-age across *any* period. That is asking too much of L&A data.

Given the way the AMA is built, its CVs ought to appropriately reflect the information content of the CKMR data itself, but not necessarily the additional— and very important— information content from the L&A data. Overall, I suspect (and within this study I can do no more than suspect) that the predicted CVs for different sampling strategies over the same period are reasonable at least in relative terms, and are not unreasonable in absolute terms over the the medium-term: say, until 2020. Before then, it will be necessary to revisit sampling design anyway, as the currently-untested assumptions are clarified by new data, and the objectives of CKMR itself in CCSBT have become more defined.

[13]Whereas the CKMA assumed no *expected* trend 2003-2010, though random fluctuations from year-to-year could still induce a *realized* trend.

Appendix C. Genotyping methods for CKMR

C.1. **Background.** Finding POPs in a wild population is no longer an especially difficult task with either microsatellites or SNPs, notwithstanding that great care is required. Bravington et al., 2014 showed that the 25 microsatellites developed for SBT are sufficient to apply the "exclusion principle" to determine POP status, with enough margin for error to bound the likely extent of false negatives and false positives. With SNPs, a lot more loci would be needed, but the ideas are similar. Similar calculations indicate that 300 well-chosen SNPs should be sufficient (see Box 1). Finding HSPs is much more challenging, and will require many more SNPs; so if we can genotype well enough to find HSPs, there will certainly be no problem with finding POPs. We therefore concentrate on the HSP case below.

In future, we will have larger sample sizes of SBT, which will mean many more comparisons because of the quadratic dependence of comparisons on samples (all adults versus all juveniles). However, the number of comparisons will not increase indefinitely, because the demographic chance of POP eventually becomes negligible when "juvenile birth" and "adult recapture" are separated by very long intervals. Eventually it becomes pointless to do very-distant comparisons, because the (minuscule) risk of false-positives outweighs the (by then even more minuscule) chance of finding anything informative. Suppose we expect eventually to make comparisons across a maximum birth-recapture window of 20 years, i.e. about 5X longer than to date, and that annual sample sizes are comparable. This would lead to about 10^9 comparisons in any one birth-recapture window. How many SNPs might be needed to ensure a safe gap between near-false-positives in Unrelated Pairs, and false-negatives from genotyping error in true POPs?

With a MALF of 0.5 (the best case for exclusion), the per-locus probability of exclusion (two different homozygotes) in an UP is 1/8. With 300 such SNPs and 10^9 comparisons (almost all being UPs), only one UP would be expected to have fewer than 8 apparently-excluding loci. Assuming a quite pessimistic per-locus genotyping error rate of 3%, then only 4% of true POPs would be expected to have 8 or more apparently-excluding loci through errors (see below). Thus, by deciding to call a POP if and only if the number of apparent exclusions is fewer than 8, we could exclude false-positives almost entirely (1 expected in maybe 1000 true POPs, extrapolating from current results) while sacrificing only a small proportion of true POPs above the threshold. The expected proportion sacrificed can be estimated based on the observed error rates, for which there will be plenty of data among the true POPs. For abundance-estimation, this amounts to having a known "tag loss rate", which is easily incorporated into the mark-recapture setting. This more formal approach is a little different to our previous approach with microsatellites, which was simply to show that any errors from false-positives or false-negatives would be small compared to the overall CV. In fact, though, if we use enough SNPs to find HSPs, there will be no ambiguity at all about POPs, so the simpler approach will be fine.

In fact, this calculation overestimates the number of loci required, because exclusion is not the only source of information: AA/AA and BB/BB pairs are more common in POPs than in UPs, so for marginal cases close to the cut-off (8 above) there should be clear information to make the decision. Alternatively, one could opt directly for a likelihood-based rather than pure-exclusion-based decision rule; the caveat with POP-finding is that an explicit allowance for genotyping error is required, otherwise a single apparent exclusion means zero likelihood of a POP.

To justify the 4% figure: suppose the inherited allele for a POP at a locus is A (the other possibility being B). The only way a single genotyping error at that locus can lead to false exclusion, is if (i) the non-coinherited allele is truly A in the first animal and B in the second, and (ii) the second animal is wrongly genotyped as B at the coinherited allele, so that an AA/BB combination is seen. Of the four copies in an AA/AB POP, there is only one at which a single genotyping error would yield a false exclusion, and because of (i) only half the pairs will be AA/AB in the first place. Therefore the chance of a false exclusion in a POP with MALF of 0.5, is half the per-locus error rate, i.e. here 1.5% by assumption. The rest is just Binomial probability calculations.

Box 1: Exclusion calculations with SNPs

To find HSPs in SBT in a statistically reliable way, it turns out that we will need to score reliably "quite a lot" of SNPs (say, 1500-2000); later, we will show why this number is necessary, and that it is indeed achievable. The number is much larger than the number of chromosomes ($2n = 46 - -48$), so that *linkage* is certain to occur; in other words, if You and Your kin coinherit one SNP, then the pair of You are also very likely to have coinherited many nearby SNPs. The SBT genome has not been sequenced, so we have no idea which SNPs will be linked, simply the guarantee that many of them will be. However, 1500-2000 SNPs is not high enough to give much risk

of *linkage disequilibrium*; in other words, for any individual, there should be almost no correlation between its genotypes at any two different SNPs (or, rather, that there are only a handful of SNPs-pairs which happen to be physically close enough on the genome to exhibit such correlation). For theoretical and empirical backup to that claim, see section C.2.2.

Once HSPs have been found, the descent (maternal or paternal) can be inferred by comparing the mtDNA of the pair; this must be the same if they share a mother, but will (usually) be different if they share a father.

C.2. **Strategy for finding HSPs.** As with our previous POP-only CKMR study, we take a two-stage approach, first making a yes-no judgement on whether each pair (of juveniles) is a HSP, and then just using those binary outcomes in the abundance-estimation step. The sensitivity of conclusions to any possible errors in the yes-no judgements are assessed post hoc. In principle, one could instead consider allowing for ambiguity about each individual pair's HSP status into the abundance-estimation model, so that the two steps are done jointly; however, we see that very much as a last resort, to be followed only if one is forced to use inadequate genotype data for some reason. From experience, it is quite hard enough to construct, fit, and diagnose a CKMR model with yes-no pairwise "data"[14]. The only modification to "exact yes-no" that may become necessary with HSPs, is that we may forced to sacrifice (reject) a proportion of true HSPs where the relationship is insufficiently certain, in order to avoid much more dangerous contamination by "near-HSPs" that are actually unrelated. The proportion sacrificed can be estimated post hoc, and incorporated into the abundance-estimation step as a known tag-loss parameter.

The structure of our approach to HSP-finding is explained below and in Figure C.1. It covers design of genetics, design of the study itself, and post hoc analyses when enough samples have been genotyped. We assume that SNPs will be used as just described, and that every locus will be genotyped, albeit possibly with error. (This is unlike the usat case, where for safety's sake an appreciable proportion of loci had to be discarded as unscorable because of ambiguity in the graphical summaries, even in fish with generally good DNA.)

(1) Estimate population allele frequencies (can be safely assumed accurate, since there will be 10000+ samples).
(2) We need a pairwise statistic for animals i and j that has good power to distinguish HSPs from UPs.
   (a) Start with the log-likelihood-ratio (LOD) for HSP:UP at each locus ℓ, i.e. for genotypes $g_{\ell i}, g_{\ell j}$ at that locus

$$\mathrm{LOD}_{\ell ij} = \log\left(\mathbb{P}\left[g_{\ell i}, g_{\ell j}|\mathrm{HSP}\right]/\mathbb{P}\left[g_{\ell i}, g_{\ell j}|\mathrm{UP}\right]\right)$$

   computed from the allele frequencies and assuming HWE.
   (b) Define the overall statistic for the pair (i, j) to be the pseudo-LOD (PLOD), formed by summing the per-locus LODs: $\mathrm{PLOD}_{ij} = \sum_\ell \mathrm{LOD}_{\ell ij}$. The "pseudo" is because the per-locus LODs aren't independent when the true relationship is HSP, because of linkage.
(3) Compute the "null distribution" of PLOD, i.e. its distribution when the true relationship is UP: $\mathbb{P}\left[\mathrm{PLOD} < x|\mathrm{UP}\right]$ for all x. This can be calculated theoretically from the population allele frequencies, because if LD is not an issue then by definition the per-locus PLODs *are* independent under UP. In particular, the mean e_{U} and variance v_{U} are easy to compute.
(4) Given the number of comparisons in the study, and some idea of the demographics, compute:
   (a) the likely number of true HSPs, and
   (b) a threshold PLOD ξ such that the expected number of UPs with PLOD $> \xi$ is a small fraction (say, well under 1%) of the true HSPs. (See later for details)
(5) The mean PLOD|HSP, e_{H}, is also readily calculated from allele frequencies, since it does not require independence between loci.
   (a) Make sure that e_{H} is well above ξ given the number of loci being considered.

[14] As one example: if each individual pair needs to be considered during each likelihood computation, it becomes impossible to group "alike" comparisons, and runtime would increase by a factor of 1000 or so.

(6) The variance $\mathbb{V}[\text{PLOD}|\text{HSP}]$ is *not* calculable in advance, because it depends on the extent of linkage. However, once enough samples have been genotyped to reveal plenty of HSPs, we can estimate $\mathbb{V}[\text{PLOD}|\text{HSP}]$ just from those pairs whose PLOD exceeds $\mathbb{E}[\text{PLOD}|\text{HSP}]$ since, by virtue of #5a, they are almost certain all to be HSPs and their mean is known. In other words, we can estimate

$$\mathbb{V}[\text{PLOD}|\text{HSP}] \approx \sum_{ij:\text{PLOD}_{ij}>e_{\text{H}}} \left(\text{PLOD}_{ij} - e_{\text{H}}\right)^2$$

(7) Given $\left(e_{\text{H}}, v_{\text{H}}, \xi\right)$, compute $p_\xi \triangleq \mathbb{P}[\text{PLOD} < \xi|\text{HSP}]$— this is the expected proportion of true HSPs that will be rejected because their PLOD happens to fall below the threshold. This can then be accommodated in the abundance-estimation step, simply as a known "tag loss rate" ν, so that there is no overall bias. For abundance estimation, the prior probability that a comparison between i and j will be *classified* as HSP, becomes this:

$$\mathbb{P}[\text{classified HSP}_{ij}|...] = \mathbb{P}[\text{true HSP}_{ij}|...](1-\nu) + (1 - \mathbb{P}[\text{true HSP}_{ij}|...])\,\nu$$

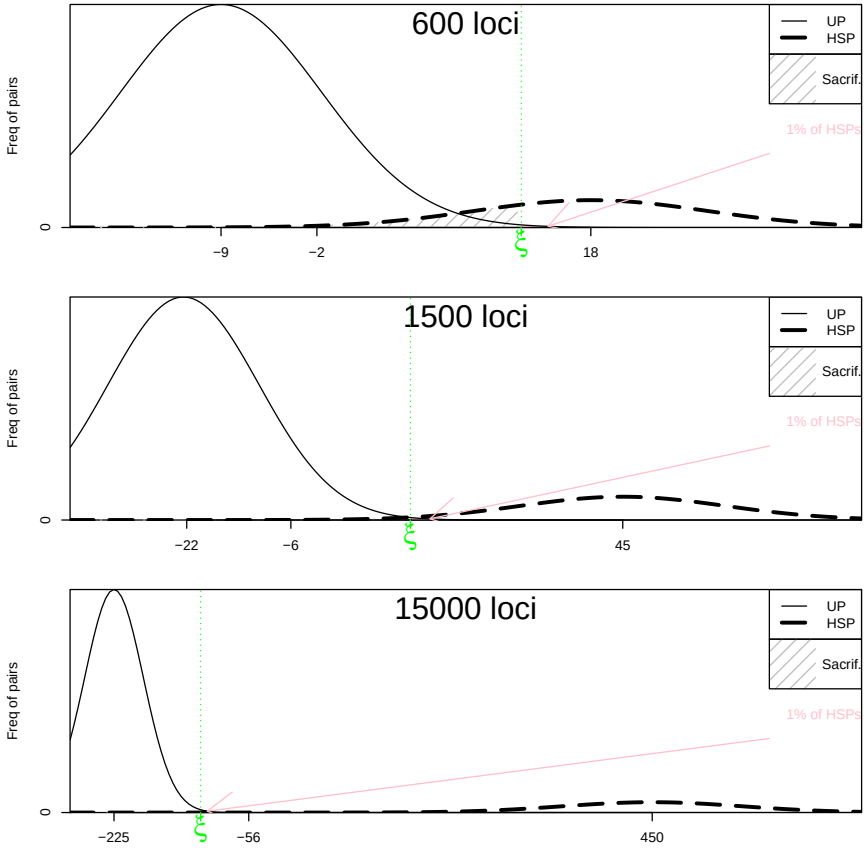
(8) The probability that any non-HS close-kin pair will have PLOD $> \xi$ should also be checked[15]. If there is more risk of false-positive HSPs from non-HSP kin than from UPs, then ξ should be raised accordingly, and step #7 should be repeated. The calculations depend on linkage, so have to be done post hoc.
Clearly, each non-HS CK pair does have a higher chance of PLOD $> \xi$ than a UP because of coinheritance, but UPs are so numerically dominant that the PLOD|UP distribution may nevertheless completely cover the PLOD|(other) distribution. We have gone through this exercise for another species (white shark), and there the PLOD distribution for first-half-cousins (shared grandparent) did protrude slightly beyond PLOD|UP, so we did have to raise ξ slightly. The same may happen for SBT, though UPs are much commoner there because of the larger population size.

(9) There is no practical way[16] genetically to distinguish HSPs from Grandparent-Grandchild Pairs (GGP), but for SBT this is irrelevant since we will only be comparing juveniles (immatures) to other juveniles.

[15] EG "first half-cousin pairs" C1P, which share one grandparent.

[16] This is theoretically possible, but only with the aid of extremely dense SNPs and a complete genome assembly.

FIGURE C.1. Finding HSPs: distributions of test statistic across pairs, each for different numbers of fictitious SNPs. UP becomes narrower and further away from HSP; latter becomes slightly narrower but only up to a point because of linkage. With even more loci, LD (linkage disequilibrium) would eventually place a "minimal narrowness" on the UP distribution.



This procedure deliberately treats UPs and HSPs quite asymmetrically. False positives with PLOD $> \xi$ (pairs classified as HSP that are really UPs) are to be excluded by design, but false negatives are to be expected and allowed for. Because UPs are so dominant— by a factor of one million or so— it is imperative to be sure that false positives do not leak through to any appreciable extent; otherwise, getting things slightly wrong (by using too few loci, or too low a ξ) could easily swamp the true HSPs in a way that cannot be accurately quantified. It is much better to be conservative in setting a high ξ to avoid bias, and to pay the price in variance caused by sacrificing a few true HSPs; as long as enough loci are used, that price will be small anyway.

There are existing software packages for finding kin, including HSPs, but we have avoided them here because they do not provide enough control over false-positive and false-negative error rates, which are *the* key issue for large-scale sparse kinship applications like SBT. There is in any case nothing to be gained statistically; when only pairs of kin are available, the above approach makes efficient use of the available information (likelihood ratios are the optimal statistics for hypothesis tests) and is quite simple to implement computationally.

C.2.1. *Genotyping errors.* The scheme above ignores the possibility of genotyping errors. In one sense, locus-by-locus errors are not as serious for HSPs as for POPs, because there is no guarantee of inheritance at any locus, so there is no rigid "exclusion principle" that can be apparently violated by an error; no one locus can really matter that much (especially since we know nothing about linkage). However, errors could still mess up the distributional assumptions and lead to bias via unexpected false negatives. It is therefore important to try to estimate errors in advance, check the robustness of the procedure against them, and adjust the procedure if required.

While certain types of genotyping error can be estimated from repeat extractions and repeat sequencings on the same sample, that does not fully the the relevant issue for HSP-finding; conceivably, some "errors" might be quite consistent within an individual but not heritable. For SBT, we do not have any known HSPs[17], but we are in the happy position of already having 45 known POPs, which can be used to estimate the *heritable error rate* based on apparent exclusions. This is a much better basis for predicting the error rate within HSPs, since the same mechanisms are involved. Details are shown in section C.11.

Large-scale errors (contamination, rotated plates, mislabelled samples) of course require different treatment. Contamination is more of a concern with SBG than because the techniques are so sensitive to minuscule amounts of DNA; we have seen this with preliminary runs on both SBT and shark samples. Contaminated samples are easily detected when large numbers of SNPs are used, via an unusually high proportion of heterozygotes. Other large-scale errors need protocol safeguards such as plate-specific blanks, controls, etc., which are important in any study; our previous microsatellite work on SBT showed it can be done, so we do not repeat the details here.

C.2.2. *Will Linkage Disequilibrium matter?* The scheme above assumes that we can compute the distribution of PLOD for UPs just from knowing allele frequencies, which is only true if the loci are uncorrelated, i.e. in Linkage Disequilibrium (LD). This would certainly cease to be true if the number of loci was extremely large, but the number required for HSPs is modest compared to many studies.

In other wild populations, the spacing distance between loci required for LD to decay varies widely. For example:

- From Gray et al., 2009: *"In 11 gray wolf populations and one coyote population, we find that the extent of LD as measured by the distance at which$r^2 = 0.2$ extends <10 kb in outbred populations to >1.7 Mb in populations that have experienced significant founder events and bottlenecks. This large range in the extent of LD parallels that observed in 18 dog breeds where the $r^2 = 0.2$ value varies from ∼20 kb to >5 Mb."*
- Backström et al., 2006 report a distance of about 400-500kb in collared flycatchers, noting LD seems *"more extensive than for other species"* (i.e. a longer distance than usual)
- Laurie et al., 2007 report *"about 100kb"* for wild mice near Tucson, Arizona. A similar figure has also been used for humans.
- Spacings of <1kb have been reported for several plants, nematodes, and small flies.

The SBT genome is about 0.7Gb(ase), so with 2000 SNPs the average distance between adjacent SNPs would be 350kb. The actual distance required for LD to decay to irrelevance depends[18] on the historical effective population size N_e, which for SBT should be very large; the stock assessment suggests that the pre-exploitation female adult abundance (census size) would have been about 10,000,000. While the link between N_{true} and N_e is notoriously weak, there are several reasons why we would expect N_e to be high for SBT: large census size; no obvious reason for bottlenecks in the past (current abundance is at an all-time low); no population substructure (only one spawning ground); long spawning season with many spawning events per individual so little prospect of any one individual's offspring dominating a cohort (although larger adults will contribute much more than smaller ones).

Regardless of what we might or might not expect from the SBT genome, the real point about LD for CKMR purposes is that it only matters insofar as it distorts the variance of the PLOD|UP distribution. In the roughly 5000 UP comparisons so far, the empirical variance is close to predictions, so there is no suggestion of serious LD (section C.12.1). Even if LD was apparent, the variance of PLOD|HP could be estimated from its empirical distribution (perhaps restricting to values below the theoretical mean for UPs) and used in place of the theoretical variance to set ξ.

[17] Actually, there is one; see later.

[18] Among other things... for example, it is not constant across the genome, but then we have no idea where the SNPs are on the SBT genome anyway.

C.3. **Choice of genetic technology.** There are two approaches for genotyping SNPs: "Targetted Assays" (TA) and "Sequencer-Based Genotyping" (SBG[19]), which can be described as follows in very non-technical and conceptual terms. Both have several different implementations, and the detail can matter.

TA: is an established approach with high reliability (very low error rates) and low unit cost if large numbers of samples but only small numbers of loci are needed [20]. TA may be the cheapest option for individual genetic mark-recapture, where only 50 well-chosen SNPs may suffice, and perhaps for known-possible-parents POP-finding in hatchery studies (e.g. Anderson and Garza, 2006). It is essentially a black-box presence-absence procedure for each allele at each locus. However, the cost of TA scales proportionally with the number of loci, in multiples of say 50 or 100; with 300 SNPs needed for POP-finding in SBT, we estimate that the minimum unit cost of TA would be similar to usats, though the automation and reliability of TA is an advantage. While the costs of TAs have fallen somewhat over the years (as have microsatellites), the rate of cheapening is not spectacular.

SBG: (including GBS, RADseq, ddRAD, nextRAD) uses a sequencer to read entire DNA fragments, physically mixed but labelled (sometimes described as "bar-coded") to distinguish which sample each fragment came from. Depending on how the fragments are obtained, there can be thousands to millions from each individual, some or all of which may contain SNPs. The sequencer reports the actual base pairs on each of the billions of fragments it examines, and software is used to stitch together these "reads" and to work out how many copies of each allele are seen in each individual. The output is not automatically a "present or not" for each locus/allele/sample, but rather a count of sequences read; this can then be processed into "present or not" by a subsequent software step, using algorithms that we discuss below.

SBG is newer, and there is a confusing range of possibilities, some of which will be much better and/or cheaper than others for large-scale pairwise kinship in CKMR. After much preliminary consideration, including tests on other species, we have settled on one particular variant: ddRAD with an intermediate step so that sequencing focusses only on a selected set of loci of interest. We refer to this here as "HFS-Dart"; the company Dart (Diversity Arrays Technology Pty Ltd; www.diversityarrays.com) has prime experience in ddRAD, and builds in a number of crucial QC steps directly. HFS-Dart offers very low unit cost (currently about half of the usat cost) and the high reliability essential for identifying not just POPs, but HSPs too; not all SBG methods would be equally reliable. The cost structure of SBG is quite different to TA or microsatellites. For individual mark-recapture where very few SNPs are required, HFS-Dart would probably be slightly more expensive than TA, since individual barcoding currently sets a lower limit; however, the unit cost only changes slightly when the number of loci is increased to 2000, which is ample for HSPs. There are other versions of SBG which might be cheaper than HFS-Dart if individual mark-recapture or wild POPs were the only target, but the overall cost basis of HFS-Dart for SBT is compelling because if HSPs can also be found then fewer samples are needed (as well as the robustness of abundance estimation being qualitatively increased). Over the past decade, all genetic processing has become cheaper, but it is sequencing which has seen the most dramatic drops in cost, and this trend is likely to continue; HFS-Dart may become even cheaper in the future.

As we show later, HFS-Dart offers several compelling advantages for large-scale close-kin:

- Low unit cost because, with irrelevant parts of the genome discarded, more individuals and loci can be sequenced in the same run. The one-off cost of development for focussing is also cheap, so that it becomes cost-effective to switch from standard Dart to HFS-Dart around the 1000-sample mark, even aside from the benefits of higher coverage (average read-depth of say >100 per copy). Those benefits are:

[19]Unhappily, the phrase "genotyping by sequencing"— which logically should cover a whole suite of related techniques— also sometimes refers to one very specific proprietary technique "GBS", and is the subject of ongoing lawsuits. To avoid problems, I've simply invented the term "SBG" to cover the whole lot.

[20]IE cost of genotyping each sample at all required loci, after sample collection and DNA extraction (which cost the same regardless of method). Other less-useful definitions of "unit cost", e.g. of one SNP per individual at low reliability, are rife in the literature.

- It is basically impossible to miss an allele, if the allele is truly present;
- It is easy to see fake alleles resulting from sequencer errors or other artefacts, because the counts from errors are clearly much lower than true alleles;
- For loci where the average read-depth reaches several hundred, it is also possible to reliably detect individual null alleles, i.e. to distinguish genotype AO from a true homozygote AA. This transforms null-alleles from being a statistical nuisance that reduces information content, into an advantage; the "3rd allele" can considerably increase the power of the locus for kinship, therefore reducing the overall number of loci needed.

C.4. **Selection of loci for focussing, and testing of HFS-Dart.** We began by running Dart's ddRAD process on one plate of 94 SBT samples, including most of the known POPs. About 30,000 SNPs were found. Null alleles were common, with many loci showing an excess of apparent homozygotes. However, most potential SNPs were not useful in kin-finding, for one of a variety of reasons:

- Minor Allele Frequency below 0.1 (much the commonest reason to avoid);
- high frequency of null-alleles;
- low average read-depth (say, under 10);
- evidence of "paralogs": the same sequence occurring in more than one place on the genome (at least in some individuals). This is manifest in an excessive proportion of "heterozygotes", since each allele gets two or more chances to appear;
- other unhealthy diagnostics in the count data, such as substantially higher average counts for one allele compared to the other among clear heterozygotes. Dart Inc. use a range of similar quality measures;

The last two are reasons for outright rejection, but the first three are less serious, especially for HFS-Dart. It turns out that the expected ΔPLOD is rather insensitive to MALF, at least for MALF$>$0.1; the AA/BB case, which is most informative about coinheritance, is rare when MALF is low, but very informative when it does occur. Also, null alleles— which immediately bring to mind complications and reduced information content— actually make a locus *more* informative *if* they can be individually scored (i.e. if we can separate AO from AA). Low read-depth can often be alleviated by lab optimizations, and the focussing step also helps even this out. The pool of potentially-suitable loci thus turned out to be rather larger than we first thought.

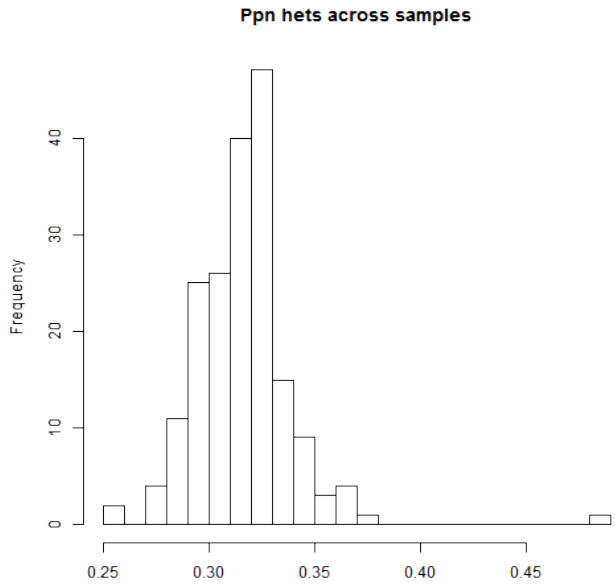
We chose an initial set of 1500 candidate SNPs (more could have been chosen by relaxing the MALF or NALF stringency). Kinference calculations showed that, with accurate genotyping[21] and given the estimated MALFs and NALFs, these 1500 should be ample for HSP-finding. The next step was to confirm genotyping accuracy using HFS-Dart. Time and budget precluded development of the focussing step for all 1500, so a subset of about 770 was randomly selected for initial focussing, of which 644 were ready in time for this analysis.

The HFS-Dart process was then applied to two new plates of SBT (188 samples). The first plate comprised some replicate samples from individuals with DNA processed in different ways, most of the known POPs, and a "Mystery Pair" of apparent non-POP close-kin found from usat study. The second plate was taken directly from an existing "template plate" used for the usat study, of juveniles from 2007. As in 2014, we checked per-sample rates of heterozygosity[22] across loci to look for contamination (mixed samples in the same well)— SBG is much more sensitive to contamination than usats or TA. In 2014, there were at least 10 such samples, physically near each other on the plate and suggesting a plate. In the 2015 plates, only one sample showed clear excess heterozygosity (Figure C.2), and it was excluded from further analysis.

[21] More accurate than possible with the read-counts from "vanilla" ddRAD.

[22] There is a chicken-and-egg issue here, in that some kind of rule is needed for calling genotypes in order to scrutinize the data and work out how best to call genotypes. We used a simple for the heterozygosity check, but the same rule was good enough to reveal some problems in the 2014 samples.

FIGURE C.2. Contamination check via ppn. heterozygous loci per sample: obvious outlier sample on RHS.

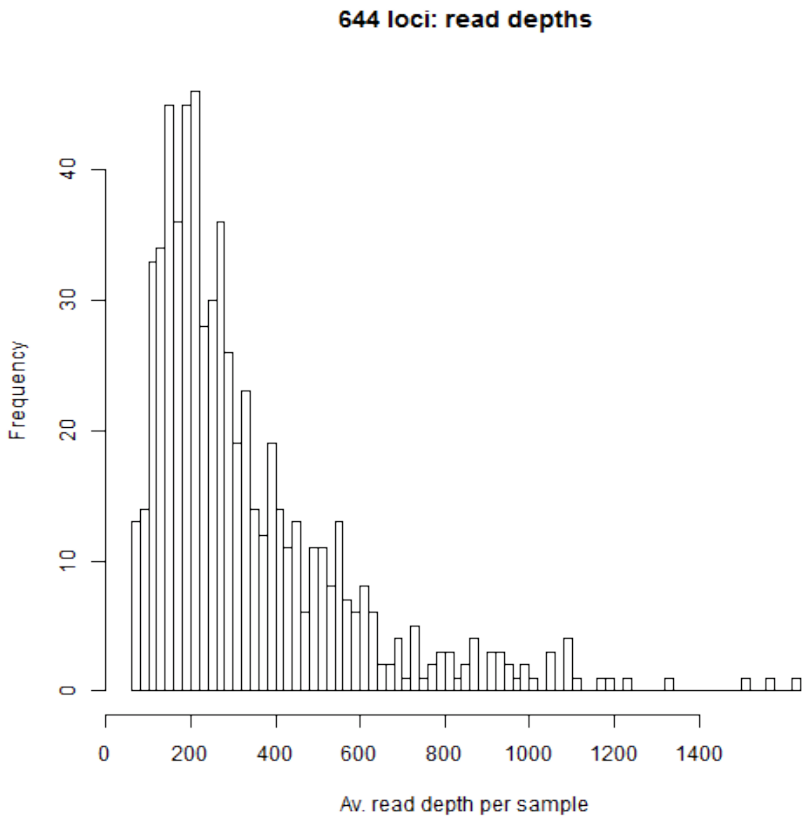


In the next few subsections, we show diagnostic results from these 644 loci, and discuss how the results would scale up to full set of 1500 or 1800 loci (which would have the same unit cost per sample).

C.5. **Read-depths.** Read-depths (counts) were normalized across samples (to allow for differences in DNA concentration) and across alleles (to allow for slightly different PCR rates— a fairly small effect), but not across loci. The frequency distribution across loci of average total count (combined alleles) is shown in Figure C.3. Average read-depth (coverage) is typically 200 or more, and about 10X higher than the "unfocussed" read-depth[23]. Even at the lower end, there is no sharp drop-off in average read depth; the lowest 5 are (55,60,65,65,66). The Figure in fact underestimates true read-depth for many loci, since it includes samples with null alleles.

[23]From the 2500 best loci in the 2014 standard ddRAD run.

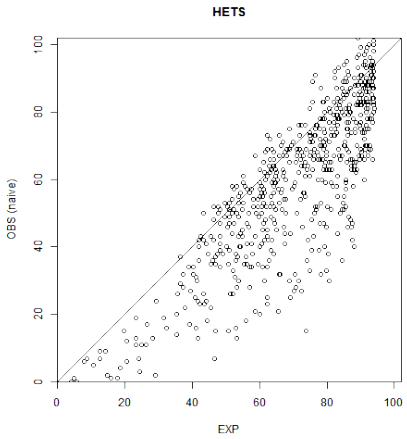
FIGURE C.3. Average read depth (including nulls)



Of course, it is not just average read-depth that matters for genotyping accuracy, but consistency too. We examine this in the next sections.

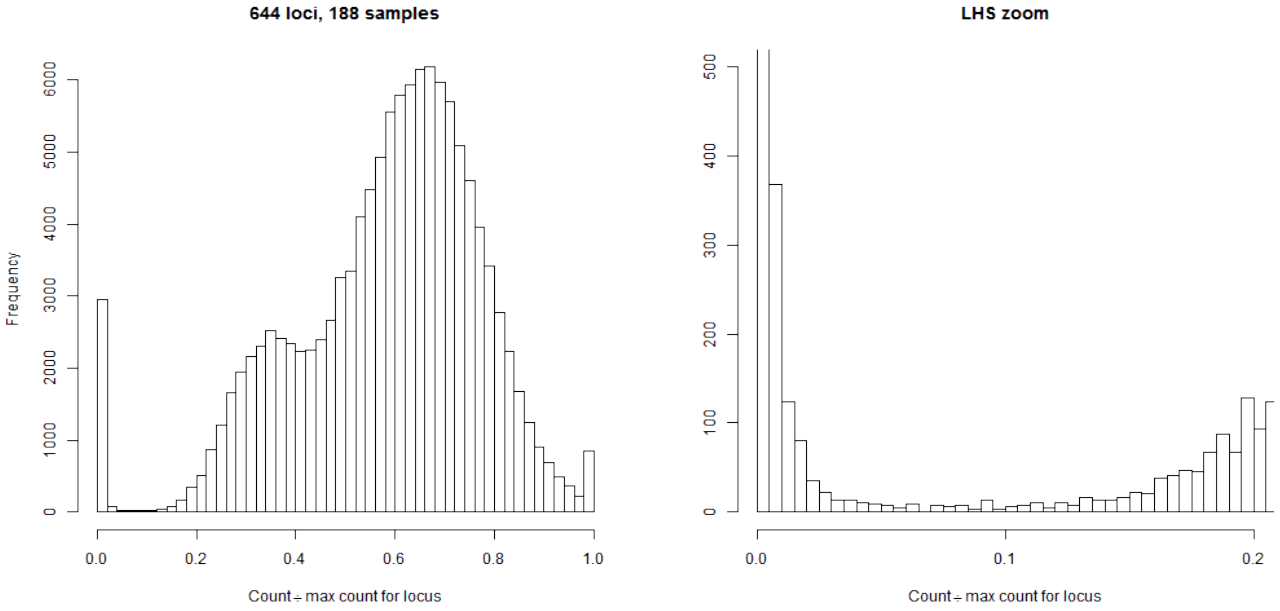
C.6. **Double-nulls.** The 2014 results showed strong departures from HWE at many loci (excess homozygosity / heterozygote deficiency), consistent with null alleles (Figure C.4). There is nothing surprising about seeing null alleles with RAD, because restriction enzymes are deliberately specific, so that a mutation at the restriction site will mean that the corresponding allele simply does not make it to the PCR stage. (ddRAD has two such enzymes, so will yield more nulls than vanilla RAD.) Focussing does not change this. Other types of mutation could also lead to bona fide nulls: indels at the restriction site, for example, or microsatellites or other insertions near the SNP site which could shift the fragment size outside the selected size range. Of course, there are other reasons why HWE can fail— stock structure being notorious— but there is no reason to think these other cases might apply to SBT, and the HFS-Dart make it possible to check directly.

FIGURE C.4. Non-HWE under a purely-two-allele model. MALF here estimated from simple two-allele model, excluding double-nulls (see below) from estimation and from expected values. The strength of the effect is slightly overemphasized because the samples include replicates and POPs, but most loci clearly show some "heterozygote deficiency".



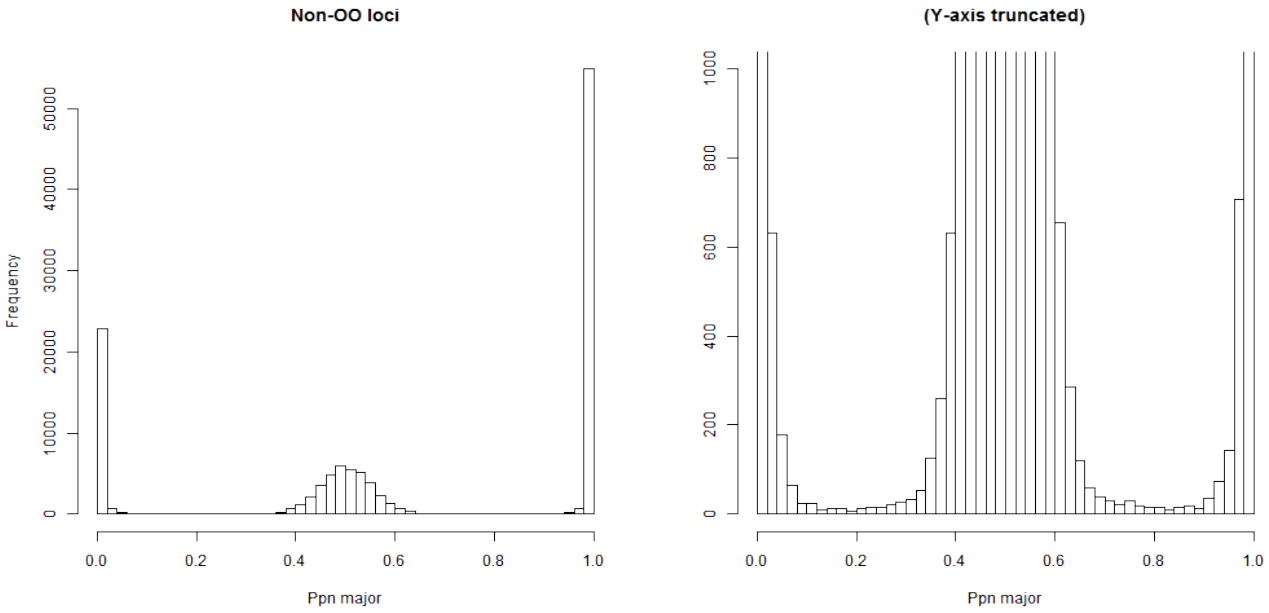
With average read-depths of at least 50, it should be straightforward to at least see double-nulls. Indeed, they are visible as a spike on the left of Figure C.5a, where counts (read-depths) are scaled to the maximum seen for that locus across all samples. Small spurious counts can occur from e.g. sequencer error, "bleed-thru" from bar-codes on different individuals, or minuscule levels of contamination, but with such high read-depths these cases are clearly distinguishable from genuine alleles. Figure C.5b is a magnified version of the LHS of Figure C.5a, and shows clear separation. Our first step in genotyping was therefore to score a locus as double-null if the total count at that locus was below 0.07 of the maximum count for the locus. (Counts per sample are normed across all loci, so that each sample has the same total count.) Figure C.5a also shows a bump around x=0.35, as expected if there are single nulls; see below. Note that double-nulls will be much rarer than single nulls.

FIGURE C.5. Double nulls



C.7. **Clarity of heterozygote/homozygote calls.** After removing double-nulls ("OO"), we can examine the proportion of major-allele reads in each locus/sample. The results show an unambiguous distinction between AB, A-no-B, and B-no-A in almost all cases (Figure C.6). Only 0.29% of loci fall into the ambiguous range $(0.1, 0.3)$ or $(0.7, 0.9)$ where misclassification might be expected. The reason for saying "A-no-B" rather than "AA", is that the apparent homozygote might actually be AO (A-null) instead; see next.

FIGURE C.6. Proportion of A (major) alleles per non-OO locus

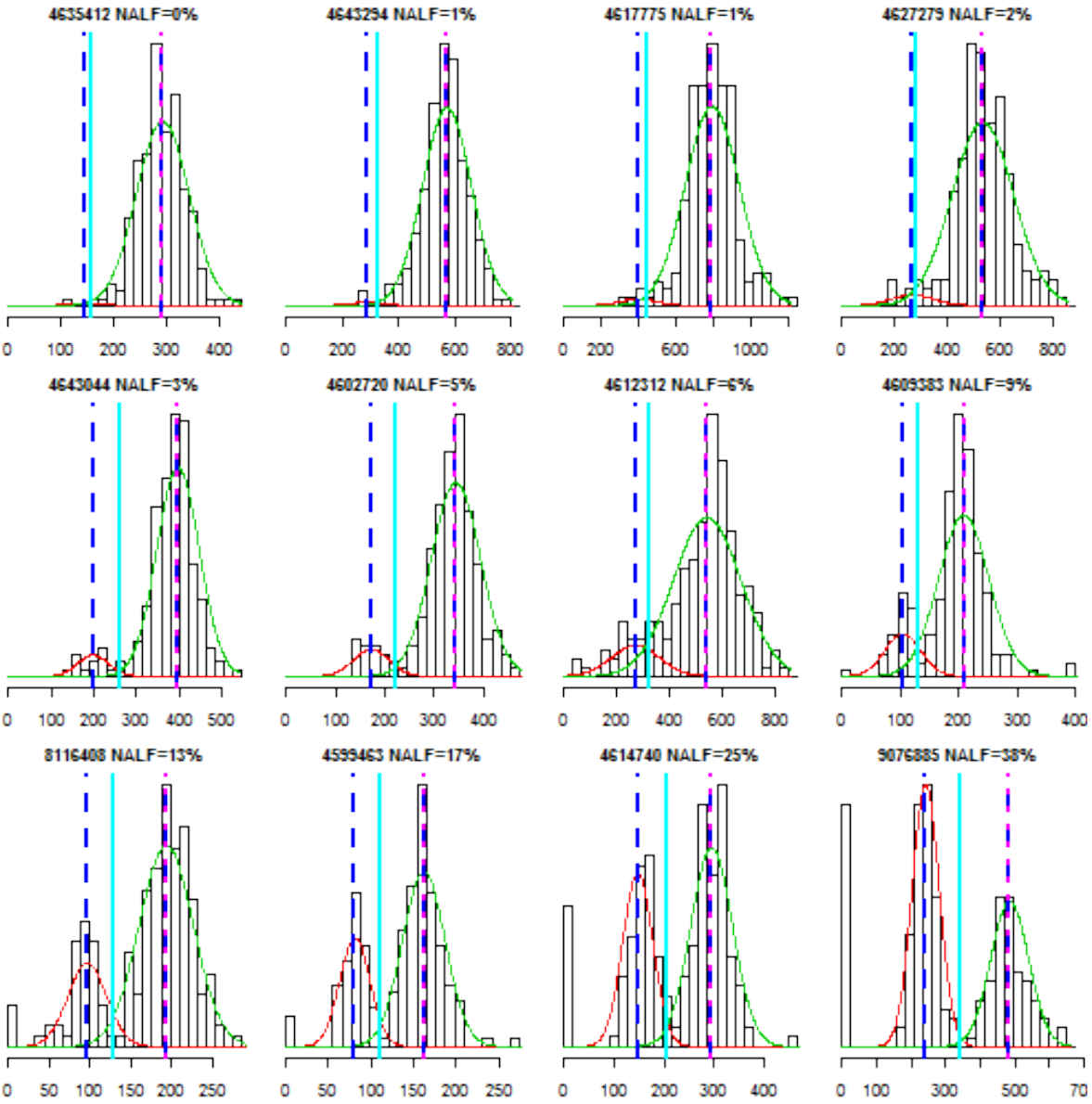


C.8. **Single nulls.** For non-OO non-heterozygote loci, the high read-depth may make it possible to distinguish between true heterozygotes (AA) and single-nulls (AO). This would be useful for kinship because it would provide a 3rd allele; the drawback of SNPs for kinship normally, is that there are so few alleles and the information per SNP is very weak, so extra alleles are desirable in principle. Whether this is useful in practice depends on the sample-to-sample variability in counts, and the classification error rate. The fall-back position for any locus is simply to call just 4 genotypes instead of 6, and allow statistically for the possibility that AA is really AO; this is less informative but not catastrophic. The option of whether to call nulls can be made locus-by-locus.

Figure C.7 shows histograms of total count for 12 loci that typify the range of null-ness (specifically: these loci fall on even-spaced quantiles of the distribution of NALF across loci, as estimated later). From about the loci with appreciable NALF, two bumps are clearly visible, and at high levels of null-ness a spike from double-nulls is visible at (almost) zero counts. To quantify the degree of separation, we fitted a two-component Normal mixture to each distribution (after excluding double-nulls), with the mean and variance of the larger component constrained to be twice the mean and variance of the smaller, as one would expect if the larger corresponds to two amplifed fragments and the smaller to just one.

FIGURE C.7. Identifying single nulls

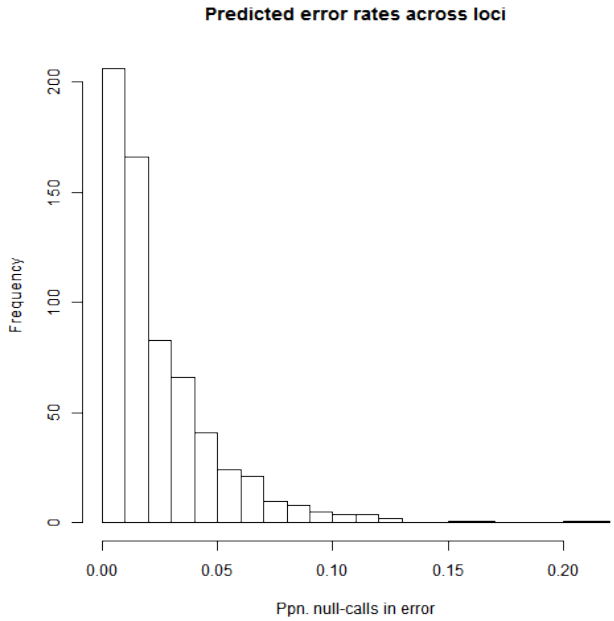
```
## [1] 0 0 0 0
## [1] 0 0 0 0
```



Frequency across samples of total counts for 12 loci; NALF shown in titles, increasing first across and then down. Dashed vertical lines are means from mixture fit; green and red curves are mixture fits; solid turquoise line is best estimated cutoff.

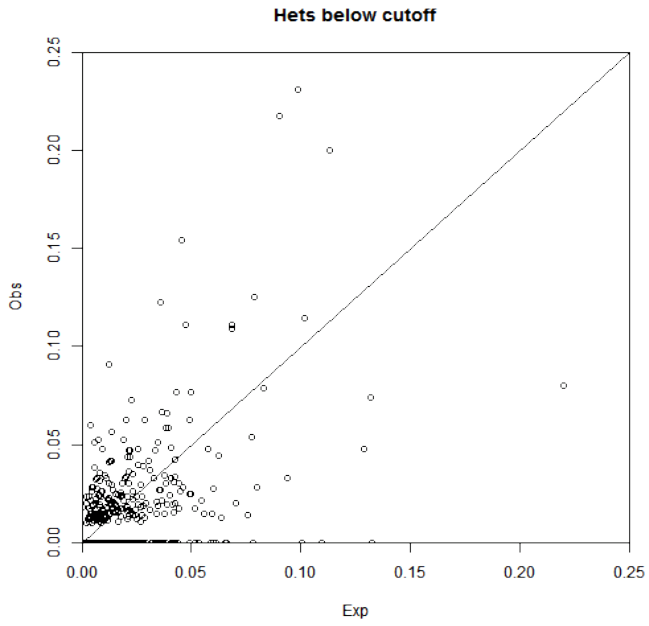
The turquoise line shows the optimal cut-point for calling presence or absence of a null, given the fitted mixtures (i.e. equal probability of "type I or II" errors). The distribution of predicted error rates across loci is shown in Figure C.8; the median is 1.61%.

FIGURE C.8. Predicted null-calling error rates across loci



These predictions assume Normality for the mixture components; a quick empirical check which avoids that assumption, is how often an AB genotype actually falls below the no-null cutoff. This corresponds very well to the predicted values (Figure C.9; slope of binomial GLM without intercept of 1.12).

FIGURE C.9. Observed and expected ppn of heterozygotes below optimal null-calling threshold



In subsequent analyses, we somewhat arbitrarily chose a threshold predicted error rate of 3% for calling nulls. In other words, if the predicted null-calling error rate is below 3%, we will call single nulls for that locus (resulting in 6 possible genotypes), and otherwise we will call "AAO" for non-heterozygotes, without attempting to decide between

AA and AO. About 29.3% of loci are deemed 4-geno, and the remainder are 6-geno. The decision sequencing for genotyping any locus/sample is:

- Double-nulls are checked first, using the criterion described earlier;
- then heterozygotes;
- for 6-geno loci, the remaining apparent homozygotes are assigned to either true homozygotes or single-nulls.

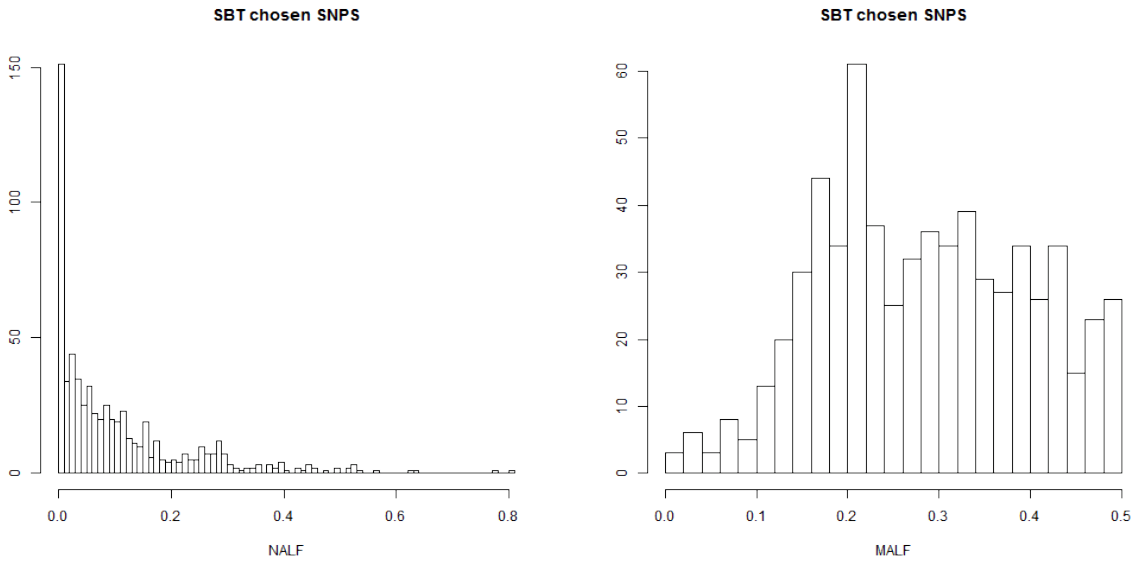
The loci in Figure C.7 were selected automatically, and their count distributions are typical of the full set. Some loci do show more overlap between the bumps than seen in Figure C.7, and a few show some clear instances of "Copy Number Variation", whereby the locus has been duplicated in some individuals. CNV, which is apparently a much more common mutation than base-changes in a SNP, could lead to mistaken calls of AO as AA, but the incidence was low overall (by visual inspection). Errors of that kind are not detectable from replicates, but the POP-based error assessment is able to estimate their frequency.

C.9. **Estimating MALF and NALF.** MALF and NALF for each were estimated simultaneously by maximum likelihood. This is trivial for 6-genotype loci, where all genotypes are observed exactly; for 4-genotype loci, numerical maximization is required.

The estimates are approximate because of the limited number of samples and the high proportion of close-kin. In a full study with 10s of 1000s of samples, allele frequency estimates will become almost exact, so this source of uncertainty is essentially irrelevant for design purposes.

The distribution of estimated NALFs and MALFs across loci is shown in Figure . The Figure somewhat overstates the incidence of nulls, because these are *point estimates* based on a limited sample of between 150-200 fish (alleles in POPs are not independent); the distribution of true NALFs would be systematically narrower, but it is not easy to correct for that with small sizes. Eventually SBT sample sizes will be so large that the small-sample issues will disappear (as it has with microsatellites).

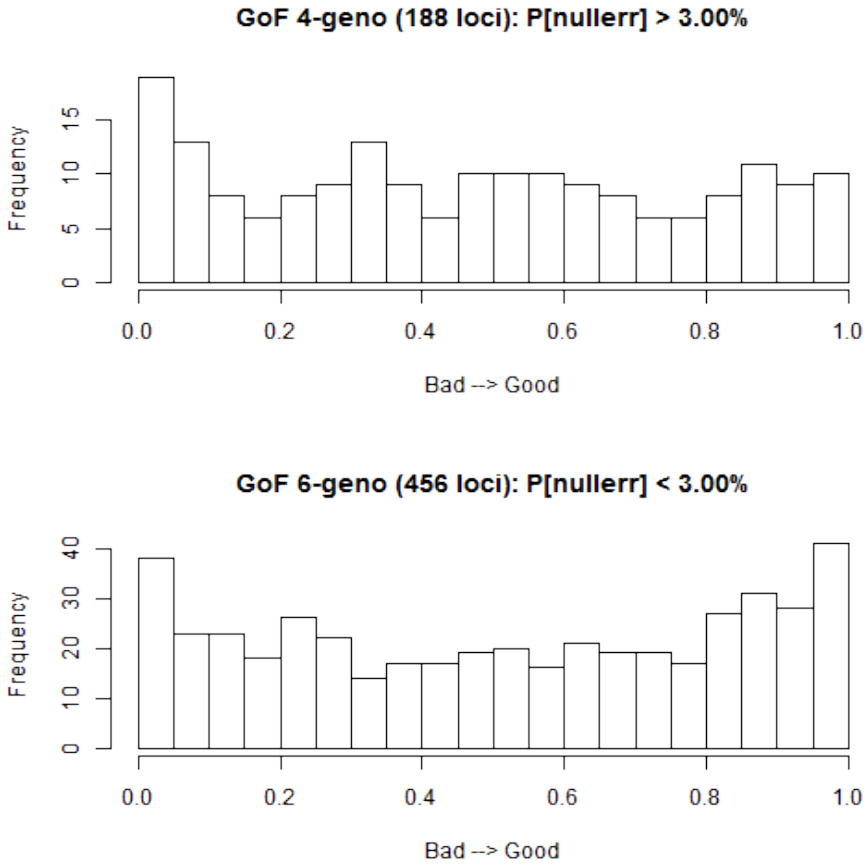
FIGURE C.10. NALF (LHS) and MALF point estimates, using 4-genotype model for all loci (i.e. not trying to call single nulls)



C.10. **Goodness-of-fit of genotypes.** With either 4 or 6 genotypes, and only two parameters that are supposed to determine their frequencies under HWE, it is possible to check goodness-of-fit by chi-squared. Figure C.11 shows

the frequency distributions across loci of the resulting p-values which, if the genotyping and HWE assumptions apply, should be uniformly distributed. And so they are[24].

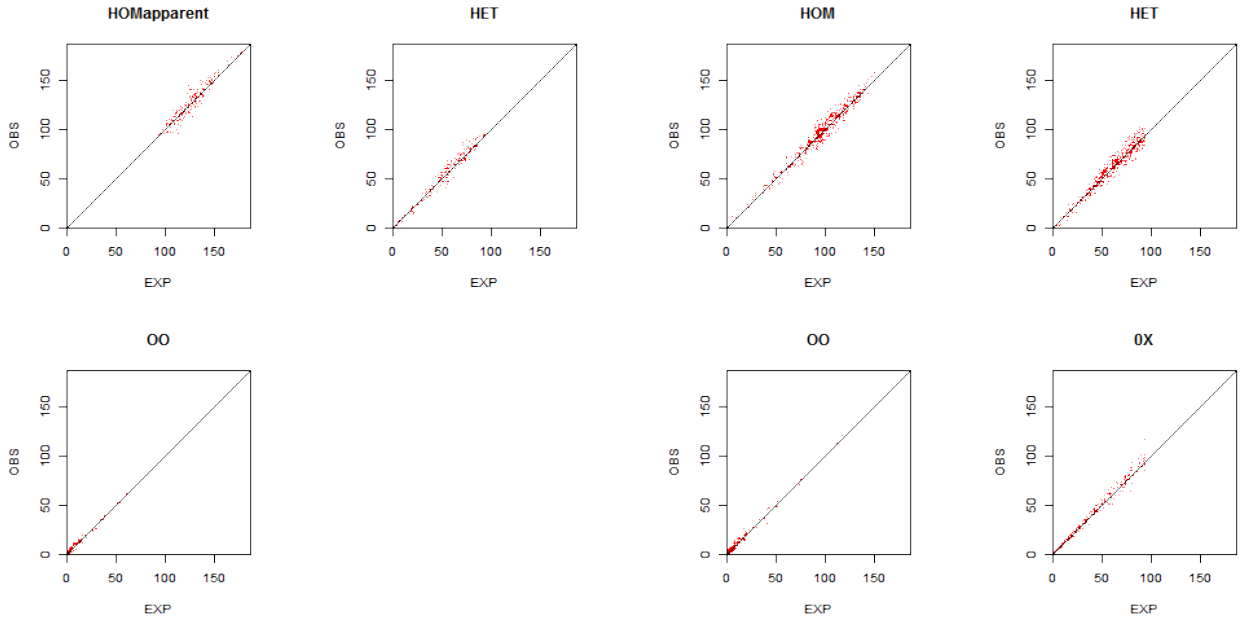
FIGURE C.11. Distribution of χ^2 p-values for GoF, for 4-geno and 6-geno loci (1 and 3 DoF respectively)



The observed and expected counts of different categories of genotypes (true homozygote, heterozygote, single-null, double-null) are shown in Figure C.12. The fits are excellent.

[24]The 194 samples include about 20 replicates, and 45 of the remainder are POPs. This would tend to exaggerate p-values slightly, stretching them towards 0.

FIGURE C.12. Observed and expected genotypes by category: 4-geno loci LHS, 6-geno loci RHS. Each dot is one sample and one locus. Expected values are calculated assuming HWE with MLEs of MALF and NALF.

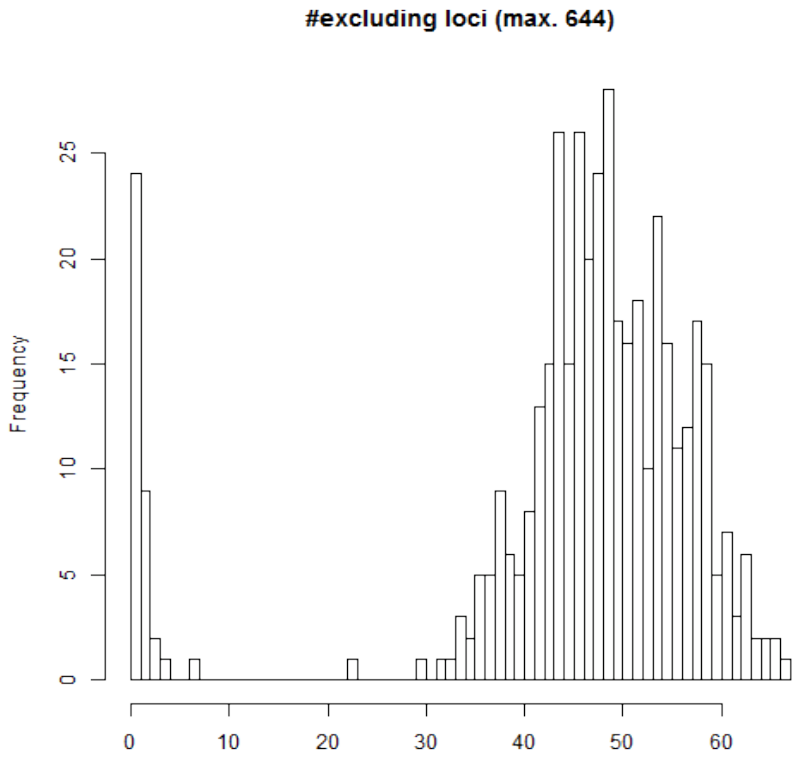


C.11. **Error rates.** Genotyping "errors" include not just "physical errors" in what the sequencer and post-processing software *thinks* it saw, but also in the algorithms we have used to assign genotypes on the basis of reported sequence counts. There are two sources of information on error rates in our HFS-Dart plates: replicate samples, and POPs. The error rate in replicate samples is extremely low: 99.87% of loci/samples have identical genotypes. Since there are four copies per genotype-comparison, the per-copy error rate in replicates is about 0.032%.

Because it is entirely possible to make the same mistake twice in a replicate, POPs provide a more reliable basis for estimating error rates insofar as they apply to inheritance and thus to finding HSPs. A simple summary is to look at the distribution of number-of-parentage-excluding loci[25] across POPs, including some non-POPs for reference (Figure C.13). The POPs and UPs are obvious, and there is an interesting Mystery Pair that excludes at 23 SNP loci. In the usat study, this pair stood out clearly from both POPs and UPs because it excluded at only 2 of 25 usat loci, suggesting that the pair might be a non-POP close-kin pair. The SNP analysis confirms that suggestion; 23 excluding loci is about half-way between POPs and a typical UP, in just the right place for Grandparent-Grandchild or Half-Sib status. (Note that this particular criterion— number of exclusions— is definitely *not* the most powerful way to find HSPs or GGPs, as already discussed. Note also that we propose using a lot more than 644 loci in the full-scale application. The separation between HSPs and UPs will be much stronger than seen here.)

[25]The definition of "exclusion" for a locus depends on whether 4 or 6 genotypes are being called; see later. Based on predicted proportion of single-null errors, % of the loci were treated as 6-genotype here.

FIGURE C.13. "Exclusion" count in POPs, some UPs, and the Mystery Pair



The breakdown of exclusion errors by type is shown in Table 3; for 4-genotype loci, confirmed exclusion (correct or not) only occurs with (AB,OO) genotype-pairs. Among the 6-genotype loci, by far the commonest error is AO/BB, which can occur with just one genotyping error among the four alleles: if the BB is really be BO; all other possibilities require at least two errors, or a misinterpreted OO. All the AB/OO errors are in 4-genotype loci, which tend to have lower read-depth.

TABLE 3. Breakdown of exclusion errors in POPs, for 6- and then 4-genotype loci. A can be either the major or minor allele, and B is always its counterpart.

	AB/OO	A0/BB	AA/BB	AA/OO	[All6]
Cases	0	26	2	7	16835

	AB/OO	[All4]
Cases	9	6993

Detailed inspection of the raw counts in these exclusions indicates that the errors are generally not subtle: e.g. a count (major/minor allele) of 156/0 in one member, and 0/319 in the other. In other words, they are not sensitive to what choices are made about thresholds. The most likely explanation for most errors is Copy-Number-Variation in a non-null allele, which (according to the 6-genotype criterion used here) can lead to the null allele being ignored. A very few loci just seem to have amplified weakly in some fish, so that double-nulls are called; Dart Pty Ltd commented that DNA quality was rather variable in the mostly-POP plate, which we assembled from a variety of original tissue samples stored for different numbers of years.

Amongst the 4-genotype loci, the numbers of AA/BB pseudo-exclusions are very consistent with the estimated NALFs, which of course do not use any information about POPs (Table 4).

TABLE 4. Numbers of AA/BB pseudo-exclusions for 4-genotype loci

	X0	X1	X2	X3	X4	X5
OBS	71.00	66.00	29.00	14.00	5.00	4.00
EXP	59.66	68.79	39.66	15.24	4.39	1.01

Since AO/BB-type exclusions are much the commonest in Table 3 and result from a single genotyping error (calling a BO as BB), we can estimate the inherited-single-null error rate ε empirically. The true genopair must be AO/BO, and the proportion of genopairs that truly have this value can be predicted from the MALF and NALF for each locus. Each time, there are two chances to make a mistake with a null (mistakes with the non-null alleles do not trigger exclusion). The *expected* number of errors for a locus across all N_{POP} pairs is therefore

$$\mathbb{E}\left[\#\mathrm{AO/BB}\right] = N_{\mathrm{POP}} \times 2 p_A p_B p_0 \times 2 \times \varepsilon$$

The error rate is clearly small enough that a Poisson approximation to the Binomial is accurate, so ε can be estimated via a Poisson GLM to give $\hat{\varepsilon}$ =4.12%. This is considerably higher than the replicate-only analysis would suggest— perhaps a cautionary tale about the limited value of replication on its own for estimating error rates. It is also slightly higher than the 3.0% "prediction threshold" for (non-inherited) single-null errors used to assign loci to 6- or 4-genotype status; in principle, we would like the 6-genotype loci to have single-null-error rates below that threshold. But the key thing is that this estimate of $\hat{\varepsilon}$ captures inheritance-related effects that could affect HSP inference, and is still a *small* number. In the next subsection, we allow for the impact of inherited-single-null errors for on HSP-finding.

Other types of error in Table 3 seem rare enough that their impact on HSP-finding can be safely neglected (there is no exclusion principle with HSPs, so no single error among 1000–2000 SNP loci can matter much).

C.12. **Reliability of HSP-finding for SBT.**

C.12.1. *Empirical distribution of PLOD when truth is UP.* We can compare the theoretical distribution of PLOD-given-UP with the empirical distribution of PLOD among the non-POP pairs in our plates. Among other things, this serves as a check on LD; strong LD would inflate the empirical variance. The results (from a random subsample of 3000 UPs) show no problem: Obs 56.10; Exp 59.51.

There is nothing to check for HSPs since (i) we do not have any (except the Mystery Pair, by accident) and (ii) we do not know in advance what $\mathbb{V}\left[\mathrm{PLOD}|\mathrm{HSP}\right]$ should be.

C.12.2. *How much will UPs and HSPs overlap?* This is based on section C.2, now with allowance for genotyping error. This is straightforward, given that we have estimated the error rate. Without error, we would compute the r^{th}moment $M_\ell^{(r)}$ of the PLOD for some locus ℓ with MALF/NALF π_ℓ, and for some kinship-type $k \in \{\mathrm{HSP}, \mathrm{UP}\}$, via

$$M_\ell^{(r)}(k) = \sum_{g_1 g_2} \left(\mathrm{PLOD}_\ell\left(g_1, g_2\right)\right)^r \times \mathbb{P}\left[g_1, g_2 | k, \pi_\ell\right]$$

where

$$\mathrm{PLOD}_\ell\left(g_1, g_2\right) \triangleq \log\left(\frac{\mathbb{P}\left[g_1, g_2 | K_{12} = \mathrm{HSP}, \pi_\ell\right]}{\mathbb{P}\left[g_1, g_2 | K_{12} = \mathrm{UP}, \pi_\ell\right]}\right)$$

i.e. weighting the possible PLODs by the probabilities of *true* genotypes (g_1, g_2). With error, we weight instead by the probabilities of *observed* genotypes. This is done via what is sometimes called a "confusion matrix" containing

the probabilities of observing a genopair given the true genopair:

$$M_\ell^{(r)}(k,\varepsilon) = \sum_{g_1 g_2} \left(\mathrm{PLOD}_\ell(g_1,g_2)\right)^r \times \left(\sum_{g_1^{\mathrm{true}} g_2^{\mathrm{true}}} \mathbb{P}\left[g_1,g_2|g_1^{\mathrm{true}},g_2^{\mathrm{true}};\varepsilon\right] \times \mathbb{P}\left[g_1^{\mathrm{true}},g_2^{\mathrm{true}}|k,\pi_\ell\right] \right)$$

Mean and variance for $k = \mathrm{UP}$ are obtained by summing mean and variance across loci (since loci are independent under UP in the absence of LD). For $k = \mathrm{HSP}$, the mean can be summed in the same way, but the variance remains unknown. To scale up the results from 644 loci to some proposed number such as 1500, we simply scale (i) the mean and variance of total PLOD|UP (since the loci are independent for UPs in the absence of LD), and (ii) the mean of PLOD|HSP.

To identify HSPs, we will eventually need to set a threshold PLOD ξ. Here we consider two potential thresholds, rather arbitrarily set as the 1/2-way and 3/4-way marks between the expected PLODs for UPs and HSPs. Table 5 shows the expected numbers of UPs above each threshold if 10^9 comparisons are made[26] per "time block"— a number which would put the true number of HSPs well into the 100s. Clearly, 644 loci are not enough, but with an eminently-achievable 1500 loci the 3/4-way line would be absolutely safe, and even the 1/2-way line would be fine if we can achieve 1800 loci (which would not affect unit cost). Genotype-calling does matter: if we avoided trying to call single nulls and only call 4 genotypes, then we would need many more loci. (The "All3" row shows the effect of not calling nulls at all, i.e. treating double-nulls as "not scored"; it is less powerful than the 4-genotype approach.) The impact of genotyping errors is small, at least in absolute terms (bottom two rows of the Table).

Some proportion of HSPs may be lost below ξ, in effect decreasing the "sample size". We cannot yet predict how many because we do not yet know the variance of PLOD when the true kinship is HSP[27]. However, the 1/2-way and 3/4-way thresholds are a long way below the mean for HSP, which we do know; thus we will surely be able to choose *some* ξ which safely eliminates UPs and which is well below the HSP mean, so the proportion of HSP lost will be much less than 50%. The statistical omens look sufficiently promising that I have simply ignored HSP-loss in the CV calculations later; even if 10% of HSPs ultimately do need to be sacrificed to keep UPs out, there will be little impact on overall CV of abundance etc estimates.

TABLE 5. Numbers of UPs likely to overflow into "HSP territory" with 10^9 comparisons: for different total numbers of loci (Nl), genotyping rules (G4or6; "AsNow" based on 3% cutoff), allowance for errors, and potential PLOD thresholds.

	Nl	G4or6	Errors	L-1/2	L-3/4
1	1500.00	AsNow	Y	14.98	0.00
2	1500.00	All3	Y	1395.90	0.00
3	1500.00	All4	Y	1009.19	0.00
4	1500.00	All6	Y	1.12	0.00
5	644.00	AsNow	Y	141090.70	25.66
6	1800.00	AsNow	Y	0.64	0.00
7	1800.00	AsNow	N	0.24	0.00

[26]These are calculated using Normal tail probabilities, which are not going to be completely accurate. We have also developed a saddlepoint approximation which will be much more accurate and will be used in practice with real datasets, but have not implemented it here for design purposes; it is not going to change a really minuscule number into a much bigger one, and the point here is just to show that HSPs are not going to be completely swamped by UPs.

[27]As a not-very-reliable reference: in the case of white sharks, the variance of PLOD|HSP seems to be about 3X higher than the variance of PLOD|UP. If translated to SBT would put about 96% of HSPs above the 3/4-way line, and 99.98% above the 1/2-way line. However, we have used fewer loci for white sharks than SBT, we only called 4 genotypes rather than 6 (no focussing was used), and the white shark genome is much larger, so it is not clear that this has any relevance to SBT.

On the whole, these predictions will tend to be pessimistic about the ability to identify HSPs, for a couple of reasons. First, I have used an arbitrarily-chosen 6/4-genotype cutoff of 3%. Since the impact of inheritance-single-null errors seems to be low at this level, the results could probably be improved by relaxing the cutoff to move more loci into the 6-genotype group. Second, about 10% of the 644 loci actually contain a 2nd SNP at non-trivial MALF. I could not use them in these calculations because they show very strong LD with the main SNP (unsurprisingly, since the separation is just 10s of bases). However, the very closeness of the SNPs means that the reads are "phased", i.e. we can read the two SNPs together from the *same* strand of DNA; this potentially makes the pair of SNPs more powerful for kinship, like a microsatellite. The computation becomes more complicated because there are more genotypes, so I did not try to do it here, but there can certainly be a useful gain in "kin power" at no extra cost from these 2nd SNPs.

C.13. **Summary: reliability of HFS-Dart for finding HSPs in SBT.**

- Genotyping with HFS-Dart seems to be very reliable. Given DNA of consistent quality and without contamination, average read depths are high, and read depths are quite even across samples.
- For most HFS-Dart loci, it is possible to distinguish single-null genotypes from true homozygotes with low error. This substantially improves the per-locus power to find kin, and makes it possible to find HSPs with a modest total number of loci.
- The genetic results from POP comparisons seem very consistent with expectations, given the properties of the method and HWE assumptions for SBT. For kinship purposes, error rates estimated from known POPs are more useful than from individual replicates, because the vagaries of inheritance are automatically incorporated.
- Finding POPs is completely straightforward once enough loci are used to find HSPs. A POP-only approach would actually be more expensive: unit costs would not be much lower because of the fixed costs of sampling, extraction, and barcoding, and a lot more samples are needed. In any case, the extra robustness to demographic assumptions and independent checks that HSPs provide are well worthwhile.
- We could not directly test for HSPs in this pilot project, because we would need to genotype several thousand juveniles to have a reasonable expectation of finding a few. However, we have been able to show clearly that the proposed technology will work. And we have already demonstrated that it does work, for two shark species— with fewer loci and with a less powerful genotyping approach, albeit with much lower numbers of unrelated comparisons.
- So far we have only tested a limited set of loci from the 1500 originally identified, but scaling up to the full 1500 or even 1800 should be no problem; the set we tested were randomly-chosen and thus representative.

References

Anderson EC and Garza JC (2006). "The Power of Single-Nucleotide Polymorphisms for Large-Scale Parentage Inference". *Genetics* 172.4, pp. 2567–2582.

Aranda G, Abascal FJ, Varela JL, and Medina A (2013). "Spawning behaviour and post-spawning migration patterns of Atlantic Bluefin Tuna (*Thunnus thynnus*) ascertained from Satellite Archival Tags". *PloS one* 8.10, e76445.

Backström N, Qvarnström A, Gustafsson L, and Ellegren H (2006). "Levels of linkage disequilibrium in a wild bird population". *Biology Letters* 2.3, pp. 435–438.

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

CCSBT (2012). *CCSBT Report of the Seventeenth Meeting of the Extended Scientific Committee*. Tech. rep. Commission for the Conservation of Southern Bluefin Tuna.

– (2013). *CCSBT Report of the Eighteenth Meeting of the Extended Scientific Committee*. Tech. rep. Commission for the Conservation of Southern Bluefin Tuna.

Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, Zhu L, Ostrander EA, and Wayne RK (2009). "Linkage disequilibrium and demographic history of wild and domestic canids". *Genetics* 181.4, pp. 1493–1505.

Hillary R, Preece A, Davies C, Bravington M, Eveson P, and Basson M (2012). *Initial exploration of options for inclusion of the close-kin data into the SBT operating model*. Scientific Committee report (OMMP workshop) CCSBT-ESC/1208/21. Commission for the Conservation of Southern Bluefin Tuna.

Laurie CC, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, Dean MD, Smith KL, Schadt EE, and Nachman MW (2007). "Linkage disequilibrium in wild mice". *PLoS Genet* 3.8, e144.

WANG J (2009). "A new method for estimating effective population sizes from a single sample of multilocus genotypes". *Molecular Ecology* 18.10, pp. 2148–2164.