REVIEWS OF CCSBT-ESC/1510/19: "SBT CKMR: OPTIONS FOR THE MEDIUM TERM"

Eric Anderson, Robin Waples, and Mark Bravington

Introduction

Mark Bravington, CSIRO, August 2015

Close-Kin Mark-Recapture relies on specialized genetic techniques and novel statistical models that are unfamiliar to most people in CCSBT ESC. This makes it difficult for the ESC to fully evaluate proposals. At the request of several ESC members, we therefore contacted two international experts[1] to request reviews of CCSBT-ESC/1510/19, focussing in particular on our recommendations for changing to a newer genotyping method, and the potential for using Half-Sibling-Pairs (HSPs) as well as Parent-Offspring-Pairs (POPs) in CKMR. In brief, both reviewers support our proposals; if the project goes ahead, then we will continue to liaise with the reviewers on points of methodological detail. The reviews, which include brief autobiographies, are attached.

My only, minor, response to the reviews is to clarify one point about document 19. Both reviewers note that an estimate of absolute abundance using total numbers of HSPs could be biassed if there is unmodelled heterogeneity in reproductive output over time ("uhroot"). That is true, but in fact this year's models in document 19 deliberately avoid using total HSP this way; in effect, only *patterns* in HSPs are used, to help estimate relative (rather than absolute) abundance as well as other parameters. The "absoluteness" of abundance still comes only from POPs, which are not uhroot-susceptible. In particular, the CVs behind the sample size recommendations should be robust to any uhroot[2]. [This is a change from last year's approach, in CCSBT-ESC/1409/44; I have altered the models this year in response to a suggestion from the ESC Expert Panel, and one consequence is to make them robust to uhroot.] All this is implicitly noted in the final paragraph of both reviews, but document 19 itself does not make the point very clearly.

[1] A third expert declined due to other commitments.

[2] Some sensitivity tests do use HSPs for absolute abundance, under the *assumption* of no uhroot bias: scenarios 6,7,9 in Table 1/2017, and 11 in Table 1/2015. Those CVs were not used for recommending sample sizes.

UNITED STATES DEPARTMENT OF COMMERCE
National Atmospheric and Oceanic Administration
National Marine Fisheries Service
Fisheries Ecology Division
Southwest Fisheries Science Center
110 Shaffer Road
Santa Cruz, CA 95060

August 13, 2015

I am a member of the Molecular Ecology and Genetic Analysis Team at the Southwest Fisheries Science Center lab in Santa Cruz. I was asked to comment on "SBT Close-Kin Mark-Recapture: options for the medium term, Contract report for CCSBT, 2015" by Mark Bravington, Paige Eveson, Peter Grewe, Campbell Davies. In particular, a focus on the proposal to switch from microsatellites to sequence-based genotyping was requested.

The lab I belong to has been heavily involved in developing the genetic technologies and resources to transition from microsatellites to single-nucleotide polymorphisms (SNPs) for genetic work on Pacific salmon. I can confirm that the advantages of using sequence-based markers (like SNPs), identified by the report, have all been realized in our work. Genotyping with SNPs has proven more reliable, faster, and cheaper than genotyping with microsatellites. Additionally, since sequence-based genotyping is more easily standardized between labs, we have been able to generate data collaboratively with multiple labs without the time-consuming and costly standardization process required of using microsatellites.

Much of our work has involved using SNPs for relationship inference—in particular, inferring pairs of parents for individual offspring. When both parents are available, about 100 SNPs suffice to identify the parent pairs accurately (Anderson & Garza 2006). Accordingly, we have had great success using Targetted Assay (TA) approaches for genotyping. As noted in the report, however, when many more SNPs are required, TA approaches become too costly. I concur with this and will add that in projects requiring 300 or more SNPs (as is required to accurately identify mother-offspring or father-offspring pairs) many salmon labs are now currently using or developing sequence-based genotyping strategies to reduce cost. These strategies include either sequencing of PCR-amplified segments ("GT-seq") or restriction-associated digests (RAD) followed by a capture step (the "Rapture" approach developed by a pioneering lab at UC Davis). The HFS-Dart approach, as I understand it, belongs amongst this broad family of methods and appears sound.

The report authors have been diligent in presenting arguments that HFS-Dart will work, and their proposed analysis strategy appears sound and has been tested as effectively as possible given the extent of their preliminary data to date. I was particularly interested in their proposed approach to use read-depth distributions for inferring genotypes in the presence null alleles. While inspection of Mendelian incompatibilities within known parent-offspring pairs showed that the genotyping error rates when calling genotypes with null alleles was higher than hoped for, these rates should not be too high to deal with. The distribution of

the number of loci with incompatible genotype calls within parent-offspring pairs shows that it will still be trivial to distinguish POPs from UPs and, I concur that inference of HSPs should be considerably less affected by genotyping errors.

The proposed analyses for inferring HSPs are sound. When using 1500 markers, many of the loci will be coinherited during any particular meiosis. While this physical linkage (in the absence of linkage disequilibrium) turns out to be irrelevant when dealing with POPs and UPs, for HSPs it will increase the variance of the log-likelihood ratio statistics used to score different pairs. In a human genetic context, today, such linkage would be dealt with by using known positions of the markers and estimated probabilities of recombination between adjacent marker pairs to calculate the likelihood ratios to take the physical linkage into account. The necessary genetic map information for such an approach is not available for SBT (as is the case for many organisms). Accordingly, the report authors propose an approach that lets them use the observed likelihood ratios for very confident HSPs to obtain an estimate of the increase in variance due to physical linkage. This allows a more refined estimate of the number of HSPs that were not detected because their likelihood ratio fell below a certain threshold. This is a nice, innovative approach to a problem that will become widespread as dense genomic data becomes more available.

The authors point out that there should be more power for inferring half-siblings by using multiple SNPs within each HFS-Dart segment. I agree completely. Our lab is currently involved in just that type of work using next-generation amplicon sequencing in a marine fish. We have been targeting 350 base-pair fragments, and occasionally find up to 4 or 5 polymorphisms (and more haplotypes/alleles) in these segments. To call short-read haplotypes requires some tweaking of the bioinformatic pipelines we use, but the increase in power offered by having multiple alleles at each locus could be quite dramatic, especially for HSP identification, as sharing of a rare allele, when it occurs, is more indicative of identity by descent than sharing of a common allele. Handling multiallelic markers will complicate the calculations and computations somewhat, but would not be infeasible. It is good to know that the assessment of power for HSP detection given in the report is hence an underestimate of what could be obtained if all the polymorphisms in a segment were used.

Finally, I agree with the other reviewer, Robin Waples, that estimating N, instead of N_e, using HSPs might rely on a few assumptions that will be difficult to assess. However, it was clear in the report that even if the N estimated by HSPs is, in the model, distinct from the N estimated by POPs, there is still an advantage to using the HSPs. Regardless, it appears the switch to HFS-Dart should save costs over microsatellites in the medium term.

Best regards,

Eric C. Anderson

eric.anderson@noaa.gov
ph. 831-420-3983

Comments on "SBT Close-Kin Mark-Recapture: options for the medium term," by Bravington et al. 2015.

I was part of the steering committee for the original study and was asked to review the above report, which describes work proposed for the next several years. We were asked to comment specifically on the cost effectiveness of switching from microsatellites to SNPs.

In general, the large number of alleles per locus for microsatellites provides some advantages for close-kin studies. However, for large-scale studies like this one the number of markers required to produce results with desired precision and accuracy can become so large that cumulative problems caused by genotypic errors become substantial. Furthermore, the rapid decline in sequencing costs in recent years means that approaches using SNPs often are more cost-effective. I believe the rationale for switching to SNPs in future SBTCK work is sound. With regard to choosing among different SNP technologies, I should acknowledge that I am not an expert in the fine details of collecting genomics data, which is a complicated process. However, the arguments presented to support the choice of using the HFS-Dart technology appear reasonable to me and I don't have any reason to suggest a different approach.

Apart from the switch in technology, the major difference in the proposed future work is to supplement the search for POPs with attempts to identify half-sibling pairs (HSPs). This is an interesting idea that also presents a number of new analytical challenges and will require additional power and more markers. First, it should be noted that although methods now exist to identify hundreds of thousands or perhaps even millions of SNPs in non-model species, these markers still have to be packaged into a relatively small number of chromosomes or linkage groups. This means that the markers are not all independent, which creates two additional challenges: 1) in some analyses it is important to quantify and account for the amount of linkage disequilibrium (LD); 2) precision does not increase as fast with addition of new markers as it would if all were independent. What this means is that, even with 10^6 or more markers, it will not be possible to resolve all levels of relatedness based only on genetic data without a known pedigree. However, half-sibs on average should share 25% of their genes, and this level of relationship is strong enough that it should be detectable using the proposed experimental design. As noted in the report, it likely will not be possible to identify all HSPs with a high degree of certainty, but it should be possible to identify sufficient numbers to provide useful information.

The main advantages of adding detection of HSPs to the program are 1) to provide more robust tests of some of the assumptions made in the POP analyses, and 2) to provide an independent estimate of N. Objective #2 is tricky because of an inherent difference between the POP and HSP analyses with respect to the entity that they provide an estimate for: census size (N) or effective population size (Ne). POPs provide an estimate of N, and the Ne/N ratio only affects the variance of the estimate (with low Ne/N you get or miss a lot of POPs at once depending on which parent you happen to sample). On the other hand, the distribution of HSPs is very sensitive to Ne and the Ne/N ratio. In fact, that is the whole premise of the sibship method for estimating Ne (Wang 2009). Wang's method is basically a discrete-generation model and assumes the samples are all from a single cohort. For iteroparous species like SBT, this means that the distribution of full and half sibs in a single-cohort sample provides an estimate of Nb =

the effective number of breeders in one year. In that case, you can't estimate N without knowing the Nb/N ratio.

The proposal seeks to get around this problem by only focusing on half-sib pairs that are from different cohorts (and hence different ages). This part gets tricky because the probability of getting half-sibs of different ages depends not only on age-specific vital rates (for which estimates are available), but also on other factors related to lifetime variance in reproductive success. These include a) Crow's "index of variability" = Vk/kbar = the ratio of the variance to the mean reproductive success of individuals of the same age and sex, and b) whether an individual's reproductive success in one year is independent of its reproductive success in another year. In species with very high fecundity, a few successful individuals can potentially produce most of a cohort, leading to very high Vk/kbar and very low Ne/N and Nb/N ratios (this is Hedgecock's sweepstakes hypothesis for reproductive success). Regarding correlations in reproductive success across years, the simplest assumption is that these events are independent (Felsenstein 1971; Hill 1972; Waples et al. 2011), but both positive and negative correlations are biologically plausible. Lee et al. (2011) evaluated the consequences for Ne of individuals that are persistently good or bad at producing offspring, and Waples and Antao (2014) evaluated the consequences for Ne and Nb of skip breeding, where energetic costs of reproduction negatively affect the likelihood of reproducing in one or more subsequent years.

The extent to which any of these apply to SBT remains to be determined, so at this point I think it is premature to conclude whether it will be feasible to obtain a valid estimate of N from HSP data. In any case, however, I expect that the cumulative information gained from inclusion of HSPs in the analyses will improve the overall performance of the method.

Robin Waples
Senior Scientist
NOAA Fisheries, Seattle WA
August 11, 2015

Felsenstein, J. 1971 Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* 68, 581–597.

Hill, W.G. 1972 Effective size of population with overlapping generations. *Theor. Pop. Biol.* 3, 278–289.

Lee, A. M., S. Engen, and B.-E. Sæther, 2011 The influence of persistent individual differences and age at maturity on effective population size. Proc. R. Soc. Lond. Ser. B Biol. Sci. 278: 3303-3312.

Wang, J., 2009 A new method for estimating effective population sizes from a single sample of multilocus genotypes. Mol. Ecol. 18: 2148-2164.

Waples, R.S., C. Do, and J. Chopelet. 2011. Calculating N_e and N_e/N in age-structured populations: a hybrid Felsenstein-Hill approach. Ecology 92:1513-1522.

Waples, R.S., and T. Antao. 2014. Intermittent breeding and constraints on litter size: consequences for effective population size per generation (N_e) and per reproductive cycle (N_b). Evolution 68:1722-1734.