# V. Large Scale Parentage Inference as an Alternative to Coded-wire Tags for Salmon Fishery Management

John Carlos Garza and Eric Anderson
Southwest Fisheries Science Center, Santa Cruz, California

Salmon fishery management and stock assessment currently use cohort-based mortality models that rely on age-specific tag recovery data as the primary annual input. The current method of choice for large-scale tagging of Pacific salmonids is the coded wire tag (CWT), a small piece of mechanically-inserted metal (1.1 x 0.25 mm long) with a numerical code that is manually cut out of the fish's head post-mortem and read under a microscope. Approximately 1 billion of these tags and nearly 600 miles of wire have been implanted in Pacific salmonids in western North America over the last 30 years.

The CWT has been enormously useful in its 30+ years of use for understanding stock composition of fishery catch, ocean distribution of different salmon stocks and age-specific mortality of all causes. It has been a crucial component of the data used by the Pacific Salmon Commission (PSC) for estimation of fishery mortality of multiple individual salmon stocks in mixed fisheries to implement the Pacific Salmon Treaty allocation of catch between the US and Canada. The management/allocation models that are used by the PSC and other management bodies (e.g. the Pacific Fishery Management Council) are cohort-based and, thus, dependent upon the cohort of origin information garnered from CWT analysis.

However, coded wire tagging and the use of the CWT data in management of Pacific Salmon Treaty fisheries currently face significant challenges. Primary among them is that CWT programs requires enormous tagging effort for a small number of tag recoveries (<2%), and is generally only applied to hatchery stocks, due to logistical problems and potential mortality associated with tagging wild juveniles. In addition, the advent of mass marking and mark selective fisheries pose serious problems to the current management system. Historically, the removal of the adipose fin clip was sequestered as an external mark for the presence of a CWT. However, U.S. law now requires adipose fin clipping of all salmon from federally-supported hatcheries. This mass marking means that up to 80% of the adipose-clipped fish sampled in some salmon fisheries do not hold CWTs. In addition, mark-selective fisheries, which are the reason for this mandate, cause a violation of one of the fundamental assumptions of the cohort analyses; that hatchery release groups experience the same mortality regimes as genetically-similar, naturally-spawning stocks. This assumption has not been adequately evaluated in Chinook salmon even in the absence of mark selective fisheries.

Genetic tagging methods have a long history in fishery genetics, having been applied to hatchery trout more than 20 years. These methods generally take one of two forms. The first is genetic stock identification (GSI) that uses genotype data and a baseline of allele frequencies to identify individuals to population/stock of origin or estimate stock proportions from a fishery mixture sample. The second is selective breeding of fish in a hatchery, such that all individuals include some unique allele or allelic combination.
However, such genetic methods have been restricted to providing population or hatchery level resolution and can not provide age of individual fish. In addition, genetic similarity between stocks/populations may limit the ability of GSI to separate stocks and selective hatchery breeding to produce genetic tags may run up against substantial operational constraints and other problems.

Several years ago, we proposed the idea of using large-scale parentage inference as an alternative to coded wire tags (Hankin et al. 2005). Originally termed *full parental genotyping*, and now referred to as *parentage-based tagging*, this genetic method does provide age of individual fish and provides exactly the same data as a CWT program, as well as significant additional information. Parentage-based tagging (PBT) is predicated upon the idea that sampling and genotyping the broodstock at a hatchery, or the spawning adults in a natural population, provides genetic tags that are recovered through parentage analysis, thereby providing a highly-efficient, transgenerational tag. For semelparous fish, the identification of parents also provides the age of that fish, not only to cohort or broodyear, but to exact date of fertilization. Since the "tagging" process requires genotyping the parents only and each female produces thousands of offspring, PBT is highly efficient, with one pair of genotypes providing thousands of tag releases. Juvenile fish are not handled at all for PBT.

The general operational routine for PBT is relatively straightforward, particularly in a hatchery setting:

1) Tissue sample broodstock adults at spawning,
2) Genotype the parental tissues with a standard set of molecular markers,
3) Create a reference "parent" database of all sampled adults,
4) Tissue sample catches from fisheries and adults in spawning escapement, and genotype these samples with same set of standard markers,
5) Query parent database to determine if parents were sampled,
6) Determine parental pair, if sampled, and thereby stock and age (cohort) of origin.

While it bears some similarity to standard GSI, PBT is fundamentally different in that it uses a type of matching algorithm to determine Mendelian compatibility of a sample with potential parental pairs present in the reference database. In contrast, GSI uses probabilistic evaluation of the alleles present in a sampled genotype to assess where its constituent alleles are present at highest frequency, as estimated from the data in the baseline database  Statistical power for GSI is highly dependent upon the number of alleles at a locus, and less so on the number of loci.  In contrast, the power for PBT is highly dependent upon the number of loci, since each locus provides an opportunity for a Mendelian incompatibility that excludes a fraction of the potential parental trios.

Such parentage analysis is a special case of the well-developed methods of pedigree reconstruction using genetic markers, which is the basis for most gene mapping and is used in legal situations to establish kin relationships. Traditionally, simple exclusion methods that rely on Mendelian incompatibilities were used in parentage analysis, but more recently maximum likelihood methods of analysis have become prevalent. However, the concept of performing parentage analysis on such a large scale, and in a mixed fishery context is novel.  Implementation of this concept required the development of additional analytical methodology and further evaluation of the feasibility of such parentage analysis when there are such a large number of potential parent pairs. We undertook such development and evaluation for large-scale parentage inference in the last two years (Anderson and Garza 2006), and established the feasibility of performing PBT for salmon management.

In this work, we determined false positive rates (the probability that a trio identified as parents/offspring was done so incorrectly) for a wide variety of potential parentage inference situations. We determined the relationship between false positive rates and the amount of genetic data necessary, as well as evaluating the effects of genotyping error and the presence of close kin in the mixture samples. We also developed two new algorithms for more efficiently evaluating potential matches in the parent database.

In early stages of this simulation study, we quickly realized that the importance of having a low genotyping error rate and a high throughput, low cost genotyping system would mean that single nucleotide polymorphism (SNP) markers would need to be the basis for any large scale application of PBT. While SNP markers are currently not widely available in great numbers for all salmon species, they will be in the next several years. Microsatellite markers can certainly be used in parentage analysis, and are currently being so employed, but it is our strong contention that they are neither feasible nor optimal for use in any coastwide application of PBT. This is because of higher genotyping error rates, the lack of portability of data and the high cost, primarily due to staff time, necessary for genotype collection. Because of the sensitivity of parentage inference to genotyping errors, and the need for large genotypes collected at minimal cost, we have determined that large SNP genotypes are the molecular marker of choice for future applications of PBT. Because of this, all of our analytical and operational evaluation of PBT has centered on the use of large SNP genotypes for parentage identification.

The evaluation of genetic data and statistical power in the Anderson and Garza (2006) work found generally that approximately 100 SNP loci would result in a false positive rate of less than one per $10^{-13}$ parent/offspring trios examined. This is a rate that is essentially without errors from the genotype data, which would make it similar qualitatively to a CWT (although the error rate for CWTs due to problems with the tag coding is not well known). This analysis assumed a genotyping error rate that was similar to the highest one reported in the literature for SNP markers (1%) and for 90% power. Trying to assign the last 10% of offspring with high confidence raises the amount of data necessary by much more than 10%. It is also worth noting that these analyses assume a mean minor allele frequency of 20%, but it will be possible to highgrade loci from among the large numbers that will be available in several years, such that the mean frequency is higher and the number of loci necessary lower.

One of the most important results of this work is the elucidation of a logarithmic relationship between the number of loci necessary for high accuracy parentage assessment and the number of potential parent/offspring trios that must be evaluated. This means that the number of SNP loci necessary for parentage analysis rises linearly as the number of parental trios possible rises exponentially. So the scope of parentage analyses necessary for coastwide implementation of a PBT program could never grow too large to be addressed with a relatively small and feasible number of genetic markers.

The presence of kin in the parent database does raise the probability of false positives for some kin relationships. What this means is that a family member may be mistaken for an actual parent in parentage inference when present. However, in general, only full siblings and double first cousins are problematic in the parentage analyses. In addition, from the point of view of cohort analysis, only false positives that incorrectly identify close kin as both parents will result in an error of importance (i.e. wrong age or hatchery). These are less likely errors than those that only misidentify one parent and therefore of less concern. Moreover, recording matings or the sorting of broodstock by date of spawning nearly eliminates the problem of false positives due to kinship.

The analyses in the Anderson and Garza (2006) study are actually quite conservative with respect to application in coastwide management of salmon fisheries. This is because those analyses were based upon the assumption that all of the fish that might need to be discriminated are part of a large undifferentiated (e.g. lacking population structure) population. When hatcheries or natural populations included in the parent database have non-zero values of $F_{ST}$ or other genetic distance measures (i.e., there is structure present), it decreases the probability of false positive parental assignments to the individuals in the parent database that are from those differentiated populations. In general, the probability of a false positive parentage assignment for an individual fish decreases by an order of magnitude with an $F_{ST}$ value of 0.05 between the population/hatchery broodstock of origin for the tagged fish and the population of origin of the potential parent. Since there is substantial population structure in Chinook salmon and no two hatcheries have broodstock with non-significant $F_{ST}$, the probability that fish from different hatcheries might be identified as close kin is even lower than found in the simulation work.

One of the concerns that arises with a PBT program is the large number of samples that might need to be genotyped for such a program, particularly if it is necessary to achieve the number of tag recoveries for smaller stocks that are currently possible with an increased CWT insertion rate and a method for external identification of fish carrying CWTs. There are several ways that the amount of genotyping data that must be collected can be minimized and the cost of implementing PBT decreased to the point of feasibility. Reducing the amount of genetic data that needs to be collected is one way to achieve this. The amount of genetic data necessary to accurately infer parent/offspring trios is dependent upon the number of potential trios that must be evaluated in the parent database. In an ideal program, or a relatively small scale one, all matings could be recorded and associated with tissue samples and genotype data. This limits the number of trios that need be examined to those including actual mated pairs. However, accurate cataloguing of all mating information and its association with tissue samples is an enormous amount of effort by hatchery staff and would not be feasible for many hatchery programs.

However, there is a useful alternative that does not require the recording of all matings but still dramatically reduces the number of parent/offspring trios that must be examined, and therefore the amount of data for accurate inference. That method is to simply separate hatchery broodstock samples by day of spawning and preferably by sex as well. This simple step, which we refer to as day bins, will decrease the number of possible trios by at least an order of magnitude and therefore the amount of genetic data necessary. Another way to reduce the amount o data necessary is to use SNP panels that have only loci with high minor allele frequencies, since a marker with two alleles at equal frequency has the most power for pedigree reconstruction. In the Anderson and Garza (2006) study, all evaluation was with marker loci that had mean minor allele frequency above 0.2 (20%). Each increase in mean minor allele frequency of 0.1 for the marker panels decreases the false positive rate by an order of magnitude. In practice, however, it may be difficult to construct SNP marker panels with mean minor allele frequency greater than about 0.3.

Other ways to decrease costs are to decrease the tagging rate by sampling and genotyping less of the broodstock. However, since PBT generally requires both parents to be sampled to achieve identifications, the decrease must be done in such a way that all sampled broodstock are from matings in which both fish are sampled. Otherwise the decrease in sampled broodstock will have a disproportionate effect on the tagging rate, since fish with only one parent genotyped will not be tagged. Perhaps the most obvious way to decrease the genotyping burden is to simply incorporate more uncertainty into the management models, either through acceptance of a higher false positive rate (i.e. more identification errors) or through smaller sample sizes from mixed fisheries.

It is hard that this point to estimate the costs of a fully implemented PBT system relative to the current CWT system. Among the reasons for this are that the costs of the CWT program, both now and in 5 years (a realistic time frame for implementation of any alternative to CWT analysis), are very hard to determine, particularly with the advent of electronic detection and mark selective fisheries. In addition, whereas CWT analysis (and microsatellite-based genetic analysis) have relatively constant costs, the costs of high-throughput SNP genotyping are decreasing rapidly as new technologies are transferred from the field of human genetics to salmon genetics laboratories. However, preliminary analyses indicate that the cost of tagging with PBT is currently lower than the cost of tagging with CWTs and that the cost of tag recovery with PBT is higher than with CWTs.

A very attractive element of a PBT program is the abundant corollary data that results from such a tagging regime. The primary additional data that comes from such a program are the many large pedigrees for multiple salmon stocks. Such pedigrees will allow determination of near parametric values for variance in family size and marine survival, and the comparison of many parameters for hatchery and wild stocks. With the successful reconstruction of large pedigrees, this project will set the stage for future estimation of heritability of physical and life history traits in Chinook salmon, which in turn will allow the prediction of the consequences of different hatchery protocols and fishery regimes. This is also the first step in the mapping and identification of the genes responsible for characters such as fecundity, age at maturity, and run-timing, which will be of great interest to both geneticists and fishery managers.

Another attractive element of a PBT program is the prospect of integration with a traditional GSI program. Such an integrated PBT/GSI program would allow fishery sampling to proceed without respect to mark or tagging status, since all fish would provide some “tag” information. Fish from hatcheries where broodstock are sampled would be identified to hatchery and cohort of origin, and all other fish would be identified to stock of origin, or used in estimating mixture proportions. While there are some important logistical considerations that would need to be addressed for such an integrated program, the prospect of a genetic sampling program where every fish is “tagged” is sufficiently compelling that it may merit further evaluation.

Among the most important logistical challenges for implementation of a coastwide PBT program of any type is the need to find standardized SNP panels that have sufficient power for parentage analysis in all indicator hatcheries. The optimization of such panels will require a larger pool of markers available from which to choose and will require a broad multi-jurisdictional effort. However, preliminary analyses of data from the human genome project, where more than 3 million SNP markers have been described, indicates that it is certainly feasible to find such a set of markers.

From the point of view of the Logistics workgroup of the Pacific Salmon Commission GSI workshops, perhaps the most important step to be taken with respect to PBT is to ensure that the elements of the multi-jurisdictional database(s) proposed be able to accommodate the data and queries that would be necessary with a large-scale PBT tagging and sampling program. This includes the very large number of genotypes that would be collected in such a program, the ability to accept very large queries that include all potential parents for a given set of tagged fish, and the ability to integrate a PBT program with the developing GSI program.

# PSC Genetic Stock Identification Workshop

May and September 2007

## Logistics Workgroup

## Final Report and Recommendations

Logistics Workgroup Members:

| | | | |
|---|---|---|---|
| Bill Ardren | USFWS | Bill Johnson | ADFG |
| John Candy | DFO | Richard Kang | NMFS, Seattle |
| Brodie Cox | WDFW | George Nandor | PSMFC |
| Doug Dehart | USFWS | Eric Schindler | ODFW |
| Carlos Garza | NMFS, Santa Cruz | Lisa Seeb | ADFG |
| Allen Grover | CDFG | Eric Volk | ADFG |
| Denise Hawkins | WDFW | Bruce White | PSC |

Logistics Workgroup Coordinator

Kenneth Johnson (Beaverton, OR)

October 22, 2007