Single nucleotide polymorphisms identified in sockeye salmon.

Report to Northern Endowment Fund : Project NF 2006 I 4

Kristina M. Miller and Terry D. Beacham

Pacific Biological Station
Department of Fisheries and Oceans
3190 Hammond Bay Road
Nanaimo, B. C.
V9T 6N7

Contents

List of Tables

List of Figures

Abstract

Effective SNP assays were developed for six genes important in migration and in-river fitness and two MHC loci. In an attempt to target "adaptive" SNPs, fifty-four genes differentially expressed in sockeye salmon during their spawning migration were used in the development of SNP assays. Some of these genes were associated with in-river fitness, while others were differentially regulated in brain or gill tissue between stocks along the migration path. The migration SNPs that were obtained were at least as informative as the MHC SNP over the range of 22 stocks surveyed (Fst's of 0.12 [MHC] to 0.16 [Gonadotropin-1 alpha], suggesting that these too could be adaptive.

Introduction

DNA contains four bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Single nucleotide polymorphisms (SNPs) are mutations in DNA that give rise to single base changes or small deletions of one or a few nucleotides. Alternately, microsatellite loci contain tandemly duplicated simple sequences of 2-6 bases, with new alleles arising rapidly through replication slippage, resulting in a gain or loss in the number of repeat sequences. When we survey SNP variation, there are only two allelic states reflecting the two potential nucleotides present at the single mutational site. Microsatellites are highly polymorphic by nature, and contain 2-100+ alleles that can be surveyed by size simultaneously. Both types of loci are capable of resolving differences between individuals and populations; however they also carry the potential to resolve different aspects of population structure. Microsatellite loci are generally found in the non-coding portion of the DNA, and are thus considered selectively neutral. As a result, the patterns of genetic variation at rapidly evolving microsatellite loci most closely reflect the demographic history of populations. Alternately, SNP variation occurs throughout the genome, including coding and non-coding regions. SNPs that occur within the coding or promoter regions of genes that code for functionally important proteins can be adaptively important as they can give rise to alterations in the function of the protein product. A prime example of adaptive SNP variation is that observed in genes within the major histocompatibility complex (MHC), whereby single base changes can alter the subset of pathogenic peptides recognized, affecting the ability of an individual to mount an effective immune response to certain pathogens (Miller et al. 2004). Hence, the survey of SNPs within genes under selection potentially enables the resolution of adaptive differentiation among populations.

The sockeye salmon stock ID baseline produced by the PBS Molecular Genetics Laboratory (MGL) currently utilized in real time fisheries management applications by the Pacific Salmon Commission contains a combination of microsatellite loci and an MHC locus, and the high accuracy and precision of this database (enabling even individual identification to stock on a coast-wide basis) is in part due to the power of combining these two marker classes (i.e. targeting demographic and adaptive differentiation among stocks). The MGL currently genotypes MHC using a mutation detection system (Denaturing Gradient Gel Electrophoresis, or DGGE; Miller et al. 1999) that has proven difficult for some laboratories to utilize. Additional MHC genes exist within the salmon genome but have not been surveyed for variation in sockeye salmon (Miller et al. 2006). Furthermore, through the gene array program on sockeye migration physiology, the

MGL has identified a number of genes that are differentially expressed among stocks during migration, as well as genes that affect the fitness outcomes of salmon during their migration in-river. Given the potential adaptive importance of these genes, they offer a prime target for SNP development for stock identification applications and for identifying levels of adaptive differentiation among stocks.

In model species, SNPs are often identified through large scale sequencing projects (e.g. the human genome), or the sequencing of large expressed sequence tag (EST) libraries (e.g. in Atlantic salmon). Unfortunately, SNPs are not generally transferable among species. Hence, SNPs have to be identified and developed individually for the species in which they are going to be applied. In less well characterized species, such as most of the Pacific salmon species, SNPs can be identified by designing PCR primers to gene sequences from closely related species, and amplifying and sequencing each gene in a range of individuals. This process is made more difficult in salmon than in most species, as the salmon genome was at one time duplicated through tetraploidization, which means that most genes contain two duplicated copies that can be difficult to differentiate. Hence, if one is not careful in identifying the sequence of both copies of a gene, many identified SNPs can arise not from allelic variation within a single gene but from variation between duplicated copies of the gene (Smith et al. 2005). Those that arise from duplicated copies of genes will be monomorphic (meaning that each individual carries both copies) and uninformative whereas allelic SNPs are polymorphic (individuals vary in the copies they inherit from their parents). Once SNPs have been identified, fluorescently-based TaqMan™ assays can be developed that will identify which SNPs are present in an individual without the need of direct sequencing of the DNA. These assays are non-gel based, and can be done simultaneous to the PCR amplification of the gene. Hence, a small number of them can be applied quite rapidly, but thus far efforts to multiplex (run multiple SNPs in the same reaction sample) SNPs have not yielded consistent, high quality results.

This project was undertaken to develop potentially powerful adaptive SNPs in sockeye salmon. While there already exists a highly effective sockeye salmon microsatellite/MHC baseline (Beacham et al. 2005), most other laboratories have not taken up the DGGE technology required to run the MHC assays, and there is some pressure to move away from microsatellites to the presumably more transportable, easy to score SNPs. However, simulation analyses have shown that it would take upwards of 200 SNPs to replace the 14 microsatellite loci currently in our sockeye salmon baseline (Beacham, Unpublished data). The cost of developing and applying that

number of SNPs is prohibitive; however, we contend that the prudent addition of a number adaptively important SNPs to a baseline containing a reduced set of easy to score, powerful microsatellite loci could result in a reduced cost of analysis of mixed fishery samples and potentially a gain in resolution among demographically related stocks.

Methods

Forty-four genes that were differentially expressed in gill and brain tissue during spawning migration of Fraser River sockeye salmon were screened for SNPS (Table 1), with nine additional genes still in development. For each gene, sequences from Atlantic salmon and Rainbow trout were obtained through BLAST searches of the EST databases from The Institute for Genomic Research (TIGR) and the Genomic Research on Atlantic Salmon Project (GRASP) and aligned in Sequencher. Because the duplication of the salmon genome through tetraploidization occurred prior to the split between Atlantic and Pacific salmon species, we attempted to identify the duplicated copies of genes through alignment of sequences from both species, with the assumption that there would only be two copies of each gene. As such, the orthologs (same gene copy) in each species tended to contain higher sequence identity than the homologs (duplicated copies) within a species. Using AlleleID 4.0 software (Premier Biosoft International), we attempted to design primers that would amplify only one of the duplicated copies of the gene, that which reflected the copy that was differentially expressed during migration. Primers were generally designed to amplify cDNA products of 500-1,000 bases. Because these were to eventually be applied to genomic DNA, which contains additional non-coding introns of variable size, we also designed some smaller primer sets to apply if genes contained large introns. In most cases, we did not have the intron-exon organization of the gene available for initial primer design. Hence, we designed roughly six primer sets per gene with the expectation that some would not, by chance, be placed over an intron.

To screen for SNPs, 16 gill and 11 liver cDNA samples from Fraser river stocks and 16 genomic DNA samples from northern stocks were amplified using the migration gene primers. For a subset of the genes with small enough product sizes (<600 bases), perpendicular denaturing gradient gel electrophoresis was performed on pooled amplified products to determine the presence or absence of SNPs therein. Alternately, if the PCR yielded the amplification of a single band of the correct size, pooled amplification products were directly sequenced, or, if this proved difficult, amplified products were cloned into a TA vector and sequenced. Sequences were obtained in both directions from each gene of interest and analyzed with SEQUENCHER 4.5

(GeneCodes Corporation). Once SNPs were identified, sequences were sent to the Assay-By-Design service from Applied Biosystems for the development of TaqMan assays.

All SNP assays were performed in 384-well reaction plates. Each 6 uL reaction contained 20 ng template DNA, 900 nM of each PCR primer, 200 nM of each probe, and 1X Taqman Universal PCR Master Mix, No AmpErase UNG (Applied Biosystems, Branchburg, New Jersey, USA). The DNA was aliquoted using a Beckman Coulter Biomek FX Liquid Handler (Fullerton, CA, USA). Each 384-well plate contained 2 no template controls, and 1 positive control for each allele. Thermal cycling was performed on an ABI 7900 real-time sequence detection system. The thermal cycling protocol was an initial 10 min incubation at 95 degrees then 50 cycles of 92 degrees for 15 sec., and an annealing-extension temperature of 60 for 1' 30 sec. Scoring of individuals was performed using the ABI 7900 Sequence Detection System 2.3 which generates scatter plots denoting which samples had which combination of alleles.

The individual SNP assays were then tested on a sample of 768 fish, including 10-90 fish from each of 22 primarily northern stocks (from Skeena, Stikine, Nass, and Owikeno drainages and SE Alaska), but also included a few more southerly stocks from the south and Central Coast of BC and the Columbia River, as well as a stock from Bristol Bay and one from Russia (Table 2) (note we surveyed twice as many samples as laid out in the proposal).

GENEPOP v3.2a (Raymond and Roussett 1997) was used to calculate Fst (Weir and Cockerham 1984), expected and observed heterozygosities, Fis and P-values for departures from Hardy-Weinburg equilibrium (HWE). Dendrograms were constructed using GDA.

Results

SNP Discovery

The genes that we targeted for SNP development were differentially regulated in gill and/or brain tissue as Fraser River sockeye salmon were migrating back to their natal streams and lakes to spawn. Some of them were important in reproductive maturation, others in osmoregulation and cellular reconfiguration in preparation for freshwater. Many were differentially expressed between stocks along the migration path. Others were associated with fitness of the salmon in-river. By targeting genes that are important in reproductive maturation, survival, and freshwater preparation, we sought to optimize our chances of identifying SNPs in genes that are adaptively

important, and therefore provide complementary information to the demographic variation we are already surveying using microsatellite loci.

Forty-four genes were surveyed for SNPs (Table 1). Of these, only 25% amplified in genomic DNA and 77% amplified in cDNA. All of the amplifications that worked for genomic DNA also worked using cDNA. When both genomic and cDNA were amplified, we sequenced products from both to find SNPs, and this approach broadened the range of individuals surveyed to include a wide sampling both in the Fraser River and Northern BC stocks. Alternately, when only cDNA amplified (likely because primers were placed over introns or products spanned large introns) our survey for SNPs included only Fraser River stocks, as this was all the cDNA we had available.

Overall, 53% of the amplified genes contained SNPS, and 55% of the SNPs that we obtained amplified in both cDNA and gDNA. Sixteen assays were designed, of which 25% were confounded by duplicated gene copies, despite the fact that we took considerable time to try to develop single gene assays. An additional 12% of the assays did not work well. Nine SNP assays were used to survey the northern populations (two were subsequently dropped due to some cross amplification among duplicated genes), while nine others are still in various stages of development. Thus far 12% of all genes and 15% of amplified genes yielded functional SNP assays. Allelic SNPS were observed at a frequency of one in every 2152 base pairs.

We had initially intended on developing a series of SNP assays on the MHC class II gene that we currently use for stock identification that would replace the gel-based DGGE assays. While the DGGE assay is quite effective at distinguishing alleles (up to 25 alleles can be distinguished simultaneously) it requires a 16 hour run, which often means that for in-season work we have the microsatellite data ready to go (in a matter of hours after obtaining the samples) and must wait until the next morning for the MHC data. In addition, there has been reluctance by other labs to take up the DGGE technology, so this gene has not been highly transferable among labs. The MHC class II exon 2 gene that we survey is 275 bases long and contains 23 variable (SNP) sites that partition between 17 alleles. As it turns out, two factors prevented us from effectively developing SNP assays to replace our DGGE assay for this gene. Firstly, there is not a high enough GC content within this fragment to develop robust TaqMan probes; this can be a problem for a wide variety of genes (i.e. just because you find a SNP does not mean you will be able to develop an assay to survey it). This is certainly a limitation to the development of SNP assays, and to the transportability of SNPs between different assay types. Secondly, due to the close

proximity of the SNP variation within this short fragment, and the fact that SNP probes and primers must match a sequence exactly in order to work properly, it would have been next to impossible and extremely expensive to develop enough probe sets to survey the range of alleles - for each SNP one could require as many as four probes to survey it. Hence, we were left with developing only a single SNP in the B1 exon, and did not extend out from this exon further, as SNPs have already been developed in other regions of this gene by Elfstrom et al. (2006). While similar problems existed in the development of SNPs for the MHC class I UBA gene, for which we also have a DGGE assay up and running, we were able to develop a lineage-detection system whereby we can fluorescently genotype each of three sequence lineages present within this gene and multiplex this assay with existing microsatellite loci (unfortunately, we did not have the genotyping data for this assay done in time for this report). Although other MHC genes are present in sockeye salmon (Miller et al. 2006), most were not polymorphic or contained duplicated genes with overlapping alleles, hence were not useful for further SNP development.

We are continuing to pursue additional migration SNPs in sockeye salmon despite the ending of this project, so that we can reach our goal of developing 15 migration SNPs.

SNP Application

Six of the seven SNPs (that were ready to go just prior to the writing of this report—still more coming) were surveyed over 22 stocks and one over 16 stocks (APOE) (Tables 2 and 3). Only one of the SNP loci, NNT_A1, displayed significant departures from HWE after correcting for multiple tests, and we suspect that this could be due to the presence of null alleles caused by additional variation under the primer or probe. Expected heterozygosities over all SNP loci ranged from 0.28 to 0.42. Virtually all stocks contained both alleles of each of the SNP loci, with the exception of three monomorphic sites for MHC class II B1 and one for APOE, but these generally contained low sample sizes (4-10 individuals that worked). To gain a measure of the range of varition among stocks, we calculated the maximum frequency difference among stocks of the subdominant allele, and obtained a range in frequencies of 0.5 (MHC) to 0.81 (NNT_A1), with a mean of 0.63. Overall Fst values ranged from 0.12 (MHC) to 0.16 (CFP1). UPGMA clustering revealed a sharp distinction McDonnell, from the Skeena River, and two SE Alaska stocks, Hetta and Shipley (Fig.1). Upon observing the allele frequency variation, it was clear that the distinction of McDonnell from all other stocks occurred at a number of SNP loci. Although some stocks clustered with others in their region, the only tight regional grouping was the Stikine, which is also very consistently differentiated with microsatellites. The southernmost out-group,

the Okanagan stock from the Columbia drainage, was not highly differentiated, and in fact clustered with stocks from the Nass and Stikine. Our most northerly stock, Kluchevka from Russia, clustered with stocks from SE Alaska and Bristol Bay.

Discussion

This SNP discovery project yielded at least seven highly informative SNPs, with more to come as we continue sequencing and testing assays. Smith et al (2005) found that SNPS are present in the salmon genome at a frequency of approximately 1 in 4,300 base pairs, whereas our more targetted approach yielded "allelic" SNPs at a frequency of 1 in 2,152 bases. The levels of heterozygosity of the SNPs in our study ranged between 24 -42%, similar to the upper levels observed in Elfstrom et al. (2006), with maximum differences between minor alleles ranging from 0.5 to 0.81. Because the SNP developed for MHC class II B1 is from a locus that overlaps with Elfstrom et al., but is a new SNP for that locus, we can compare the properties of the SNPs from both studies even though the Elfstrom study included a greater geographic range of samples (which makes it difficult to compare the Fst's directly). The three MHC class II SNPs (from a single gene) in Elfstrom's study were similar to each other in levels of heterozygosity, maximum frequency difference between minor alleles and Fst. In fact, this MHC gene was among the most informative in their study, with an Fst of 0.32-0.38 compared to an average of 0.22. Assuming that the SNP developed for this same gene in our study is equally informative (i.e. would carry a similar Fst if surveyed over the same range of samples as in the Elfstrom study), then we can compare the information content of this locus to the SNPs uncovered in the migration study. Interestingly, the migration SNPs contained even higher Fst's than the MHC gene in our study (range 0.13-0.16 compared to 0.12). Hence, our approach to targeting potentially"adaptive" SNPs from genes important in migration appears to have yielded a higher frequency of informative SNPs. The next step will be to add these SNPs to our microsatellite baseline and run simulations to find the best combination of SNPs and microsatellites to use for optimal resolution at the lowest cost for stock ID applications. With a larger baseline sample, we will also be able to identify which, if any, of these SNPs contains an "adaptive" signature, and hence will provide complementary information to the neutral microsatellite loci already in the baseline.

Difficulties Encountered in SNP Development

We acknowledge that we did not meet our targeted goal of 15 migration SNPs, although we are expecting to keep going until we do. That said, it is important to acknowledge the very

significant hurdles one must overcome when developing SNPs in salmon, as these are not necessarily difficulties present in all species.

The presence of duplicated loci, as suspected, was a significant hurdle to overcome in the design process. There are two types of gene duplications that occur within the salmon genome. The first is that generated through the whole genome duplication event that took place in an ancestral lineage that predated the salmon speciation some 20-100 million years ago. This duplication resulted in the tetraploidization of the genome. As this event was ancient, and predated the speciation of salmonids, the duplicated genes are generally fairly easy to identify (if both copies are present in the sequencing database), as the divergence among genes duplicated through tetraploidization is generally around 6-10% while the divergence within a single copy of a gene between Pacific and Atlantic salmon is only 4-5%. Hence, if you have sequence information from both duplicated copies from both Rainbow trout and Atlantic salmon, they can be readily distinguished. The second type of duplication is tandem gene duplication, and these have generally occurred much more recently, some even after speciation within the salmonids (hence species may not all carry the same complement of duplicated genes). Tandem duplications result in the formation of multiple copies of a gene located on the same chromosome, usually tightly linked. We are learning more and more through the sequencing of BAC clones that these tandemly duplicated genes are quite common in salmon, especially in the Pacific salmonids (Ben Koop, Personal Communication). The difficulty lies in the low level of divergence between tandemly duplicated genes, mostly because these events are fairly recent. Hence, it is very difficult to distinguish, based on sequence information alone, between allelic variation within one gene and variation between two tandemly duplicated genes. In this project, we "assumed" that every gene was duplicated once through tetraploidization, and carefully designed primers that would only amplify one of these copies. We then identified SNPs that were amplified from the primers specific to only one tretraploid-copy and, again, assumed for the most part that any sequence variation that was resolved was allelic. However, on occasions where more than three SNPs were found in a fragment of 500-800 bases, we were suspicious that these could represent tandemly duplicated loci, and indeed this was the case. In almost all cases where we identified allelic SNPs, there were no more than 2 variable positions (or SNPs) identified within fragments of 350-600 base pairs. On three occasions, where 7-8 SNPs were identified within a fragment (500-1300 bases), our primers appeared to have cross amplified the duplicated copies through tetraploidization. However, on 7 occasions, duplicated loci that differed in only 1-3 SNPs over 240-1100 bases were identified in fragments of bases. In all cases where duplicated loci were

amplified, the SNPs were from tandemly duplicated loci, as the resulting sequences always aligned more closely to one of the two tetraploid copies. Hence, 43% of the genes that amplified in sockeye salmon were tandemly duplicated. It is our view that this high degree of tandem duplication within the salmon genome will serve as the most significant obstacle in the future development of SNP assays.

There were other difficulties in designing SNP assays that also had to be overcome. We used EST (cDNA) sequencing databases for Atlantic salmon and rainbow trout to obtain the original gene sequences from which to design primers that would amplify genomic DNA in sockeye salmon. However, genomic DNA (gDNA) contains extra sequence (introns) between the coding exons that can vary significantly in size among genes and species. If you place a primer over the intron/exon boundary of a gene, it will amplify in cDNA (because the intron is not present) but it will not amplify in genomic DNA. Although there are programs one can use to estimate where the intron/exon boundaries are, they are not precise, and do not work well for all genes. One can also attempt to identify the structure by aligning genomic DNA sequences of other species where the intron/exon structure has been worked out, but this is very time consuming, and requires a level of expertise for each gene being developed, so in most instances was not attempted. In the end, we simply designed multiple primer sets and assumed that some would not cross intron/exon boundaries. However, we found that only 27% of the primers that amplified cDNA also amplified gDNA. While we were still able to go ahead and attempt to identify SNPs in both sets, we had a very limited range of cDNA samples (Fraser River only) with which to survey for SNP variation; we may have identified more SNPs in these genes had we had a greater range of cDNA samples. Case in point is the fact that "allelic" SNPs were identified in 50% of the genes that amplified in gDNA, and only 18% of the cDNA sequences.

As others have found, there were some SNPs for which assays could not be designed. This occurred if the SNP was too close to the end of the sequence or near an intron/exon junction or if the sequence was too AT rich for the development of high melting temperature primers and probes. Although most laboratories are currently using Taqman assays to survey SNPs, most recognize that this is among the most expensive options, as there is no capability to multiplex (run multiple SNPs at once). Although some of the salmon genetics laboratories have tried other technologies (e.g. SNPlex), most have found that 1) these assays can not be developed on all SNPs, which means if you have a combination of laboratories using different assay technolgies,

they may not be able to run all of the same SNPs, and 2) these assays are exceedingly hard to develop . In fact SNPlex, which was developed by Applied Biosystems, is no longer supported.

In conclusion, SNP development in salmon is not trivial, and takes considerable time, money and effort. Our study along with the multilaboratory efforts to develop SNPs for the development of a Chinook salmon baseline can attest to this fact. In each case, the numbers of SNPs actually developed have not met expectations due in part to issues surrounding gene duplication in salmon. Traditionally, SNPs have been developed from species that contain a sequenced genome. Using a sequenced genome, one already has genomic DNA sequence and knows how many duplicated copies of a gene exists and how they vary, which makes the development of SNP assays quite trivial. Instead, we start with no sequence information for the species we are working with, and only cDNA sequences from species that diverged 5-20 MYA. On top of this, we have very little ability to ascertain whether genes are tandemly duplicated prior to the development of SNPs. The other unfortunate fact is that unlike microsatellite loci, SNPs are not transferable among species, so one must go through this entire process of finding SNPs individually for each species. Certainly once generalized primers to amplify sequences of various genes are developed, they can be applied more quickly on a range of species, but it is still a time consuming process.

References

Beacham, T.D., J.R. Candy, B. McIntosh, C. MacConnachie, A.Tabata, K. Kaukinen, L. Deng, K. M. Miller, R. E. Withler, and N. V. Varnavskaya. 2005. Estimation of stock composition and individual identification of sockeye salmon on a Pacific Rim basis using microsatellite and major histocompatibility complex variation. Transactions of the American Fisheries Society 134: 1124-1146.

Elfstrom, CM, CT Smith and JE Seeb. 2006. Thirty-two single nucleotide polymorphism markers for high-throughput genotyping of sockeye salmon. Mol. Ecol. Notes 6(4): 1255-1260.

Miller, KM, S Li, TJ Ming, KH Kaukinen, and AD Schulze. 2006. The salmonid MHC class I: more ancient loci uncovered. Immunogenetics 58: 571-589.

Miller, KM, TJ Ming, AD Schulze, and RE Withler. 1999. Denaturing gradient gel electrophoresis (DGGE): A rapid and sensitive technique to screen nucleotide sequence variation in populations. BioTechniques 27: 1016-1030.

Miller, KM, JR Winton, AD Schulze, MK Purcell, and TJ Ming. 2004. Major histocompatibility complex loci are associated with susceptibility of Atlantic salmon to infectious hematopoeitic necrosis virus. Env. Biol. Fishes 69: 307-316.

Smith, CT, CM Elfstrom, LW Seeb, and JE Seeb. 2005. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Mol. Ecol. 14(13): 4193-203.

Figure 1. UPGMA dendrogram depicting stock structure resolved using seven SNP loci.

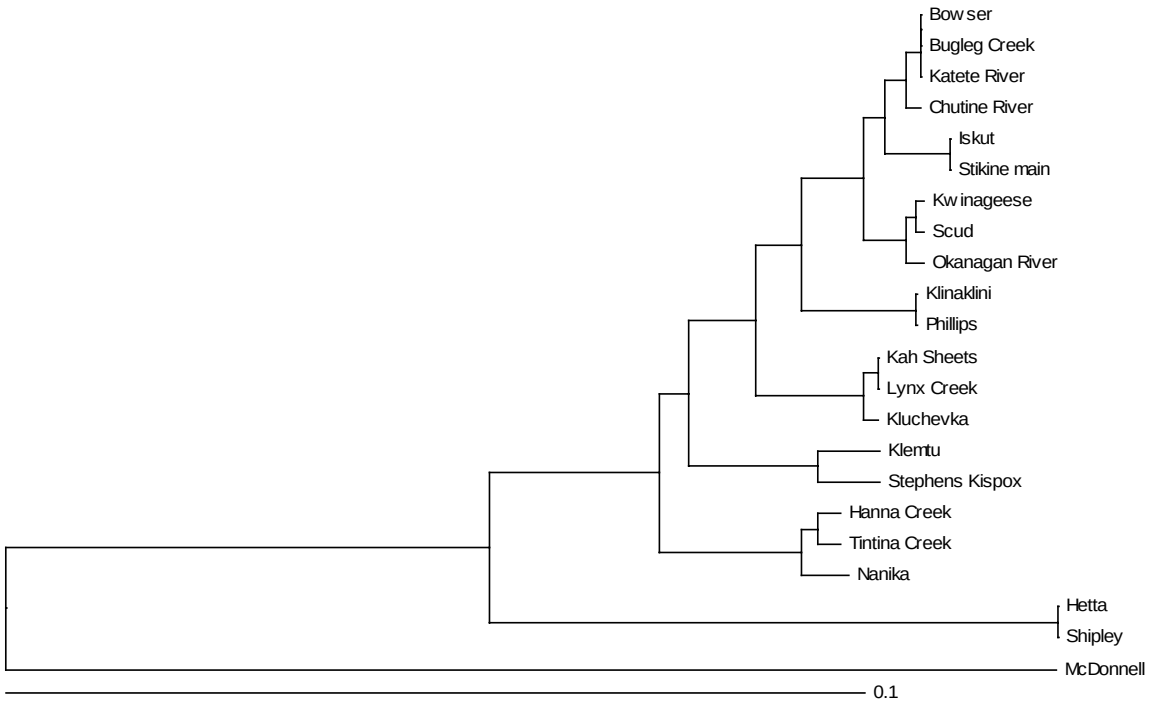

Table 1. Genes targeted for SNP ascertainment and development.

	Gene	#BP SURVEYED	cDNA amp	genomic DNA amp	SNPS	Assay	Conclusion
1	**APOE**	387	Y	Y	2	Assay By Design	good allele distinction
2	**GTHa**	592	Y	Y	1	Assay By Design	good allele distinction
3	**HSP47**	541	Y	Y	1	Assay By Design	good allele distinction
4	**CFP1**	567	Y	N	2	Assay By Design	good allele distinction
5	**NNT**	547	Y	Y	2	Assay By Design	good allele distinction
6	**PIP1**	504	Y	Y	5	Assay By Design	good allele distinction
7	**B1-p143 (MHC II)**					Allele ID 4.0	good allele distinction
8	**A2_UBA-A (MHC I)**					Allele ID 4.0	good allele distinction
9	**A2-UBA-UA (MHC I)**					Allele ID 4.0	good allele distinction
10	**A2_UBA-B (MHC I)**					Allele ID 4.0	In progress
11	**PRP**	1118	Y	N	2	Assay By Design	In progress
12	**CIRP_A**	351	Y	Y	2	Assay By Design	In progress
13	**SEPP**	663	Y	N	1	Assay By Design	In progress
14	**CTSH**	752	Y	N	1	Assay By Design	In progress
15	**CAL1B**	502	Y	N	7	Assay By Design	alleles not distinguished
16	**SAR1B**	750	Y	N	1	Assay By Design	all samples homozygous
17	**FUS**	242	Y	N	1	Assay By Design	duplicated loci not distinguished
18	**GAPDH**	1306	Y	N	7	Assay By Design	duplicated loci not distinguished
19	**RNF7-A**	612	Y	N	3	Assay By Design	duplicated loci not distinguished
20	**ATPA2**	413	Y	Y	1	Assay By Design	duplicated loci not distinguished
21	**TF-HES_A**	1151	Y	Y	2	Assay By Design	duplicated loci not distinguished
23	**ARG1**	897	Y	N	8		duplicated loci detected
24	**UCP-2**	351	Y	N	2		duplicated loci detected
25	**HbA_C**	272	Y	N			duplicated loci detected
26	**CIRP_C**	768	Y	Y	0		duplicated loci detected
27	**ATP5L**	450	Y	N	1		different copies expressed in liver and gill
22	**TF-HES_B**	1151	Y	Y	0		
28	**PHOS**	604	Y	N	0		
29	**HSP70_A**	424	Y	Y	0		
30	**COLL**	621	Y	N	0		
31	**APOEA**	502	Y	N	0		
32	**ATF4**	721	Y	N	0		
33	**CAS_Q2**	512	Y	N	0		
34	**CXCR4**	608	Y	N	0		
35	**PFKM-A**	630	Y	N	0		
36	**PFKM-B**	425	Y	N	0		
37	**HAL**	560	Y	N	0		
38	**MT1**	270	Y	N	0		
39	**MT1-B**	270	Y	N	0		
40	**PIG-F**	495	Y	N	0		
41	**RNF7-U**	500	Y	N	0		
42	**CYTOC**	110	N	N			
43	**TF-SP3**	0	N	N			
44	**DKK**	0	N	N			
45	**GH1**	0	N	N			
46	**P450**	0	N	N			
47	**DIABLO**	0	N	N			
48	**OSTM1**	0	N	N			
49	**OPI-1**	0	N	N			
50	**ANG2**	0	N	N			
51	**MP19**	0	N	N			
52	**DHCR**	in progress	Y	N			
53	**CTSB2a**	in progress	Y	N			
54	**CTSL1**	in progress	Y	N			
55	**CTSX-2**	in progress	Y	N			
56	**ISOT**	in progress	Y	N			
57	**AKID1**	in progress	Y	N			
58	**HSP90**	in progress	Y	N			

Table 2. Statistical analysis of the SNPs. Abbreviations: N=sample size, A=# alleles, He=Expected heterozygosity, Ho=Observed heterozygosity, Fis=inbreeding coefficient, Freq Range=maximum range in frequency of subdominant allele, Fst=divergence among stocks.

Protein	Abbreviation	Function	Primer and Probe Sequences	n	He	Ho	Fis	Freq Range	Fst
Apolipoprotein E	APOE	Cholesterol homeostasis	F: GTCAGATGGTTTCTCAATGGTCTCT R: ATCAAGACTCTGGAGGAGAAGCT rep 1: ACCTATAACCCCTCCATGC rep 2: ACCTATAACCCTTCATGC	277	0.395	0.365	-0.103	0.64	0.145
MHC Class I B1	B1_P143	Pathogen binding, adaptive immune response	F: GGGAGGTATGTTGGATACACTGAG R: CTTACAGACACGCTCCAGCTC rep1: TGACCCAGGATCCCAGCATCACTG rep2: TGACCCAGGAACCCAGCATCACTG	479	0.380	0.277	0.140	0.5	0.122
Cyclin fold protein 1	CFP1_1	Embryogenesis, histone modification	F: CGCAGGTCAAAGTAGTACTTAGCAT R: GAGCGTCACTTCCTGGAACTT rep1: TGCAGTTCAACATCAA rep2: CTGCAGTTCAATATCAA	728	0.416	0.372	-0.017	0.58	0.155
Gonadotrophin-1 alpha	GTHa	Reproduction	F: CAAGAAGAATCAAGAGAAAGAGAGATGGT R: CCTAGTGTCATGCACATAACGTGTA rep1: CAAGAACTAGAATGAAACAGA rep2: AAGAACTAGAATGGAACAGA	717	0.447	0.374	0.064	0.75	0.16
Heat Shock Protein 47	Hsp47	stress protein, fibrotic processes	F: CGTTCAAATAAATGCTGTTTGGCCTTT R: GTGGTGTTCGGATTTTTCCTGAAA rep1: TTATTGACTATGGCACATTG rep2: TTGACTATGGCGCATTG	723	0.456	0.377	0.006	0.513	0.146
Nicotinamide nucleotide transhydrogenase	NNT_A1	Glucose homeostasis	F: CGTATGCCAGGCCAGCTAA R: CATCTCCAGAACCACGTCGTA rep1: TCAGCCAGAAGCACGT rep2: CAGCCAGGAGCACGT	552	0.431	0.237	0.283	0.81	0.152
Solute carrier family 20 member 1 phosphate transporter	PIP_3	Osmoregulation	F: ACAGAGTCAGGACTTGATATGTACAGA R: CCTGACGAGGGTCTACTACACT rep1: AACACACATTTCTCAACACA rep2: ACACACATTTTTCAACACA	719	0.481	0.422	-0.013	0.65	0.127

Table 3. Allele Frequencies over 22 stocks.

Stock	Region	Locus	A	B
Lynx Creek	Bristol Bay_Wood River	APOE	0.97727273	0.02272727
Klemtu	Central Coast	APOE	0.63043478	0.36956522
Bowser	Nass	APOE	0.81818182	0.18181818
Kwinageese	Nass	APOE	0.79411765	0.20588235
Tintina Creek	Nass_Meziadin	APOE	0.61904762	0.38095238
Kah Sheets	SE Alaska	APOE	0.91666667	0.08333333
Shipley	SE Alaska	APOE	0.4	0.6
Mcdonnell	Skeena	APOE	0.33333333	0.66666667
Nanika	Skeena	APOE	0.5625	0.4375
Stephens Kispox	Skeena	APOE	0.67391304	0.32608696
Klinaklini	South Coast	APOE	0.73684211	0.26315789
Phillips	South Coast	APOE	0.78571429	0.21428571
Bugleg Creek	Stikine	APOE	0.86842105	0.13157895
Katete River	Stikine	APOE	1	0
Scud	Stikine	APOE	0.89285714	0.10714286
Stikine Main	Stikine	APOE	0.85714286	0.14285714
Lynx Creek	Bristol Bay_Wood River	B1_p143	0.125	0.875
Klemtu	Central Coast	B1_p143	0	1
Okanagan River	Columbia	B1_p143	0.12222222	0.87777778
Bowser	Nass	B1_p143	0.42857143	0.57142857
Kwinageese	Nass	B1_p143	0.2	0.8
Hanna Creek	Nass_Meziadin	B1_p143	0.5	0.5
Tintina Creek	Nass_Meziadin	B1_p143	0.375	0.625
Kluchevka	Russia	B1_p143	0.23529412	0.76470588
Hetta	SE Alaska	B1_p143	0.28787879	0.71212121
Kah Sheets	SE Alaska	B1_p143	0.11111111	0.88888889
Shipley	SE Alaska	B1_p143	0.42857143	0.57142857
Mcdonnell	Skeena	B1_p143	0.5	0.5
Nanika	Skeena	B1_p143	0.5	0.5
Stephens Kispox	Skeena	B1_p143	0	1
Klinaklini	South Coast	B1_p143	0.26470588	0.73529412
Phillips	South Coast	B1_p143	0.25	0.75
Bugleg Creek	Stikine	B1_p143	0.3	0.7
Chutine River	Stikine	B1_p143	0.34375	0.65625
Iskut	Stikine	B1_p143	0	1
Katete River	Stikine	B1_p143	0.5	0.5
Scud	Stikine	B1_p143	0.1	0.9
Stikine Main	Stikine	B1_p143	0.0625	0.9375
Lynx Creek	Bristol Bay_Wood River	CFP1_1	0.95652174	0.04347826
Klemtu	Central Coast	CFP1_1	0.82608696	0.17391304
Okanagan River	Columbia	CFP1_1	0.39880952	0.60119048
Bowser	Nass	CFP1_1	0.52083333	0.47916667
Kwinageese	Nass	CFP1_1	0.58695652	0.41304348
Hanna Creek	Nass_Meziadin	CFP1_1	0.80357143	0.19642857

Tintina Creek	Nass_Meziadin	CFP1_1	0.80952381	0.19047619
Kluchevka	Russia	CFP1_1	0.97058824	0.02941176
Hetta	SE Alaska	CFP1_1	0.82738095	0.17261905
Kah Sheets	SE Alaska	CFP1_1	0.97826087	0.02173913
Shipley	SE Alaska	CFP1_1	0.80645161	0.19354839
Mcdonnell	Skeena	CFP1_1	0.65909091	0.34090909
Nanika	Skeena	CFP1_1	0.375	0.625
Stephens Kispox	Skeena	CFP1_1	0.875	0.125
Klinaklini	South Coast	CFP1_1	0.7173913	0.2826087
Phillips	South Coast	CFP1_1	0.86956522	0.13043478
Bugleg Creek	Stikine	CFP1_1	0.5	0.5
Chutine River	Stikine	CFP1_1	0.56944444	0.43055556
Iskut	Stikine	CFP1_1	0.57142857	0.42857143
Katete River	Stikine	CFP1_1	0.75	0.25
Scud	Stikine	CFP1_1	0.60714286	0.39285714
Stikine Main	Stikine	CFP1_1	0.63333333	0.36666667
Lynx Creek	Bristol Bay_Wood River	GTHa	0.67391304	0.32608696
Klemtu	Central Coast	GTHa	0.26086957	0.73913043
Okanagan River	Columbia	GTHa	0.73255814	0.26744186
Bowser	Nass	GTHa	0.65217391	0.34782609
Kwinageese	Nass	GTHa	0.91304348	0.08695652
Hanna Creek	Nass_Meziadin	GTHa	0.97321429	0.02678571
Tintina Creek	Nass_Meziadin	GTHa	0.9047619	0.0952381
Kluchevka	Russia	GTHa	0.3125	0.6875
Hetta	SE Alaska	GTHa	0.54705882	0.45294118
Kah Sheets	SE Alaska	GTHa	0.58695652	0.41304348
Shipley	SE Alaska	GTHa	0.48333333	0.51666667
Mcdonnell	Skeena	GTHa	0.22727273	0.77272727
Nanika	Skeena	GTHa	0.625	0.375
Stephens Kispox	Skeena	GTHa	0.64583333	0.35416667
Klinaklini	South Coast	GTHa	0.7826087	0.2173913
Phillips	South Coast	GTHa	0.7826087	0.2173913
Bugleg Creek	Stikine	GTHa	0.725	0.275
Chutine River	Stikine	GTHa	0.68571429	0.31428571
Iskut	Stikine	GTHa	0.42857143	0.57142857
Katete River	Stikine	GTHa	0.5	0.5
Scud	Stikine	GTHa	0.76923077	0.23076923
Stikine Main	Stikine	GTHa	0.78571429	0.21428571
Lynx Creek	Bristol Bay_Wood River	HSP47	0.2173913	0.7826087
Klemtu	Central Coast	HSP47	0.43478261	0.56521739
Okanagan River	Columbia	HSP47	0.40659341	0.59340659
Bowser	Nass	HSP47	0.13043478	0.86956522
Kwinageese	Nass	HSP47	0.32608696	0.67391304
Hanna Creek	Nass_Meziadin	HSP47	0.13392857	0.86607143
Tintina Creek	Nass_Meziadin	HSP47	0.11904762	0.88095238
Kluchevka	Russia	HSP47	0.05882353	0.94117647
Hetta	SE Alaska	HSP47	0.65697674	0.34302326

Kah Sheets	SE Alaska	HSP47	0.28947368	0.71052632
Shipley	SE Alaska	HSP47	0.69230769	0.30769231
Mcdonnell	Skeena	HSP47	0.18181818	0.81818182
Nanika	Skeena	HSP47	0.25	0.75
Stephens Kispox	Skeena	HSP47	0.1875	0.8125
Klinaklini	South Coast	HSP47	0.30434783	0.69565217
Phillips	South Coast	HSP47	0.52173913	0.47826087
Bugleg Creek	Stikine	HSP47	0.3	0.7
Chutine River	Stikine	HSP47	0.31428571	0.68571429
Iskut	Stikine	HSP47	0.375	0.625
Katete River	Stikine	HSP47	0.375	0.625
Scud	Stikine	HSP47	0.17857143	0.82142857
Stikine Main	Stikine	HSP47	0.33333333	0.66666667
Lynx Creek	Bristol Bay_Wood River	NNT_a1	0.44444444	0.55555556
Klemtu	Central Coast	NNT_a1	0.19444444	0.80555556
Okanagan River	Columbia	NNT_a1	0.21875	0.78125
Bowser	Nass	NNT_a1	0.29166667	0.70833333
Kwinageese	Nass	NNT_a1	0.14285714	0.85714286
Hanna Creek	Nass_Meziadin	NNT_a1	0.27586207	0.72413793
Tintina Creek	Nass_Meziadin	NNT_a1	0.05263158	0.94736842
Kluchevka	Russia	NNT_a1	0.58333333	0.41666667
Hetta	SE Alaska	NNT_a1	0.17857143	0.82142857
Kah Sheets	SE Alaska	NNT_a1	0.58333333	0.41666667
Shipley	SE Alaska	NNT_a1	0.3	0.7
Mcdonnell	Skeena	NNT_a1	0.86842105	0.13157895
Nanika	Skeena	NNT_a1	0.07142857	0.92857143
Stephens Kispox	Skeena	NNT_a1	0.10416667	0.89583333
Klinaklini	South Coast	NNT_a1	0.59090909	0.40909091
Phillips	South Coast	NNT_a1	0.44736842	0.55263158
Bugleg Creek	Stikine	NNT_a1	0.27777778	0.72222222
Chutine River	Stikine	NNT_a1	0.57142857	0.42857143
Iskut	Stikine	NNT_a1	0.375	0.625
Katete River	Stikine	NNT_a1	0.5	0.5
Scud	Stikine	NNT_a1	0.33333333	0.66666667
Stikine Main	Stikine	NNT_a1	0.44444444	0.55555556
Lynx Creek	Bristol Bay_Wood River	PIP_3	0.54347826	0.45652174
Klemtu	Central Coast	PIP_3	0.27272727	0.72727273
Okanagan River	Columbia	PIP_3	0.49456522	0.50543478
Bowser	Nass	PIP_3	0.60416667	0.39583333
Kwinageese	Nass	PIP_3	0.52173913	0.47826087
Hanna Creek	Nass_Meziadin	PIP_3	0.63392857	0.36607143
Tintina Creek	Nass_Meziadin	PIP_3	0.5952381	0.4047619
Kluchevka	Russia	PIP_3	0.67647059	0.32352941
Hetta	SE Alaska	PIP_3	0.89411765	0.10588235
Kah Sheets	SE Alaska	PIP_3	0.675	0.325
Shipley	SE Alaska	PIP_3	0.91666667	0.08333333
Mcdonnell	Skeena	PIP_3	0.2826087	0.7173913

Nanika	Skeena	PIP_3	0.75	0.25
Stephens Kispox	Skeena	PIP_3	0.56521739	0.43478261
Klinaklini	South Coast	PIP_3	0.30434783	0.69565217
Phillips	South Coast	PIP_3	0.45454545	0.54545455
Bugleg Creek	Stikine	PIP_3	0.6	0.4
Chutine River	Stikine	PIP_3	0.55714286	0.44285714
Iskut	Stikine	PIP_3	0.64285714	0.35714286
Katete River	Stikine	PIP_3	0.625	0.375
Scud	Stikine	PIP_3	0.5	0.5
Stikine Main	Stikine	PIP_3	0.56666667	0.43333333