Development and Application of a Salmonid EST Database and cDNA Microarray: Data Mining and Interspecific Hybridization Characteristics

Matthew L. Rise,[1] Kristian R. von Schalburg,[1] Gordon D. Brown,[1] Melanie A. Mawer,[1] Robert H. Devlin,[3] Nathanael Kuipers,[1] Maura Busby,[1] Marianne Beetz-Sargent,[1] Roberto Alberto,[1] A. Ross Gibbs,[1] Peter Hunt,[1] Robert Shukin,[4] Jeffrey A. Zeznik,[4] Colleen Nelson,[4] Simon R.M. Jones,[5] Duane E. Smailus,[6] Steven J.M. Jones,[6] Jacqueline E. Schein,[6] Marco A. Marra,[6] Yaron S.N. Butterfield,[6] Jeff M. Stott,[6] Siemon H.S. Ng,[2] William S. Davidson,[2] and Ben F. Koop[1,7]

[1]Centre for Biomedical Research, University of Victoria, Victoria, British Columbia V8W 3N5 Canada; [2]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6 Canada; [3]Aquaculture Division, Fisheries and Oceans Canada, West Vancouver, British Columbia V7V 1N6 Canada; [4]Array Facility, Prostate Centre, Vancouver General Hospital, Vancouver, British Columbia V6H 3Z6 Canada; [5]Pacific Biological Station, Fisheries and Oceans Canada, Nanaimo, British Columbia V9T 6N7 Canada; [6]Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z 4E6 Canada

We report 80,388 ESTs from 23 Atlantic salmon (*Salmo salar*) cDNA libraries (61,819 ESTs), 6 rainbow trout (*Oncorhynchus mykiss*) cDNA libraries (14,544 ESTs), 2 chinook salmon (*Oncorhynchus tshawytscha*) cDNA libraries (1317 ESTs), 2 sockeye salmon (*Oncorhynchus nerka*) cDNA libraries (1243 ESTs), and 2 lake whitefish (*Coregonus clupeaformis*) cDNA libraries (1465 ESTs). The majority of these are 3′ sequences, allowing discrimination between paralogs arising from a recent genome duplication in the salmonid lineage. Sequence assembly reveals 28,710 different *S. salar*, 8981 *O. mykiss*, 1085 *O. tshawytscha*, 520 *O. nerka*, and 1176 *C. clupeaformis* putative transcripts. We annotate the submitted portion of our EST database by molecular function. Higher- and lower-molecular-weight fractions of libraries are shown to contain distinct gene sets, and higher rates of gene discovery are associated with higher-molecular weight libraries. Pyloric caecum library group annotations indicate this organ may function in redox control and as a barrier against systemic uptake of xenobiotics. A microarray is described, containing 7356 salmonid elements representing 3557 different cDNAs. Analyses of cross-species hybridizations to this cDNA microarray indicate that this resource may be used for studies involving all salmonids.

[Supplemental material is available online at http://web.uvic.ca/cbr/grasp. The sequence data from this study have been submitted to GenBank dbEST under accession nos.: *Salmo salar*, BU965588–BU965906, CA036414–CA039704, CA039711–CA064598, CA767613–CA770910, and CB498694–CB518126; *Oncorhynchus mykiss*, CB485850–CB498693; *Oncorhynchus tshawytscha*, CB484816–CB485849; *Oncorhynchus nerka*, CD510521–CD511184; and *Coregonus clupeaformis*, CB483540–CB484653. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: C. Biagi, S. Dann, S. Temple, and R. Roper stimulated the *S. salar* head kidney cells used to create one cDNA library group.]

Gene and genome duplications are thought to be primary mechanisms of increasing the number of coding sequences subject to selection, leading to new proteins, morphogenic variations, and phenotypes (Ohno 1970; Holland et al. 1994; Sidow 1996). Members of the teleost family Salmonidae, including salmon, trout, char, grayling, and whitefish, all diverged from a common ancestor that is believed to have undergone a tetraploidization event 25 to 100 million years ago, after the teleost radiation (Allendorf and Thorgaard 1984). This relatively recent putative genome duplication in the salmonid lineage is supported by karyological and genome size data. Members of the family Clupeidae (e.g., herring, alewife), thought to maintain the ancestral diploid status, have 48 to 52 mostly acrocentric chromosomes per 2N cell and genome sizes of 0.8 to 1.4 pg/N, whereas salmonids have 52 to 102 chromosomes per 2N cell (over half metacentric or submetacentric) and genome sizes of 1.9 to 3.8 pg/N (Ohno et al. 1968; Phillips and Ráb 2001; Gregory 2002). Because extant salmonids exhibit quadrivalents in meiosis (primarily in males; Ohno et al. 1965; Allendorf and Thorgaard 1984) and disomic and tetrasomic inheritance at different loci (Allendorf and Danzmann 1997), they appear to be in the process of re-establishing diploidy. Remarkably, ~50% of examined salmonid loci persist as functional duplicates (Bailey et al. 1978). Research on salmonid genomes will shed light on poorly understood evolutionary phenomena such as genome duplication and duplicate gene silencing.

In addition to their scientific importance as recent tetraploids, salmonids also serve as prominent models for studies in-

[7]Corresponding author.
E-MAIL bkoop@uvic.ca; FAX (250) 472-4075.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1687304. Article published online before print in February 2004.

14:478–490 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

volving environmental toxicology (Katchamart et al. 2002), carcinogenesis (Bailey et al. 1996), comparative immunology (Shum et al. 2001), and the molecular genetics and physiology of the stress response (Basu et al. 2002), olfaction (Zhang et al. 2001), vision (Faillace et al. 2002), osmoregulation (Tipsmarck et al. 2002), growth (Devlin et al. 2001), and gametogenesis (Madigou et al. 2002). Furthermore, Atlantic salmon (AS; *Salmo salar*) are of particular importance to the global aquaculture industry. GRASP (Genomics Research on Atlantic Salmon Project), an initiative funded by Genome Canada, is intended to improve understanding of physiological and evolutionary processes influencing the survival and phenotype of salmonids and other fish in natural and aquaculture environments. GRASP has developed genomics resources to help achieve these goals. There is a rich literature in salmonid genetics, physiology, and ecology to support these genomics research tools.

A previously reported *S. salar* EST project surveyed 1152 ESTs from six cDNA libraries, with 510 BLAST-identified sequences representing 178 salmon genes (Davey et al. 2001). There are currently (August 2003) ~60,000 *S. salar* nucleotide sequences in GenBank, of which >51,000 were submitted by GRASP. In addition to forming an EST database containing >80,000 sequences from five salmonid species, GRASP has built a microarray from 3557 unique salmonid cDNAs. Initial cross-species testing of this microarray has shown it to be effective in hybridizations with salmon, trout, and whitefish targets.

RESULTS AND DISCUSSION

EST Survey

This report describes ESTs obtained from high-complexity normalized and non-normalized, directionally cloned cDNA libraries, as well as subtracted cDNA libraries, from the following species: *S. salar* (23 libraries/library groups representing 16 adult tissues and whole juvenile), *O. mykiss* (six libraries/library groups from three adult tissues, and whole embryo and juvenile), *O. tshawytscha* (two libraries from adult mixed tissue), *O. nerka* (two libraries from adult brain and whole juvenile), and *C. clupeaformis* (two libraries from adult brain; see Table 2 below). The set of *S. salar* cDNA libraries represents most principal tissues in adult fish. EST clones are available from the corresponding author.

The 95,320 clones from these cDNA libraries (71,144 *S. salar*, 19,093 *O. mykiss*, 1824 *O. tshawytscha*, 1051 *O. nerka*, and 2208 *C. clupeaformis*) were M13 forward-sequenced and quality checked. For all libraries except SSH (suppression subtractive hybridization), M13 forward sequences of properly oriented inserts should include 3′ UTR. Because of low conservation in 3′ UTRs and the pseudotetraploidy of salmonid genomes, we focused on 3′ sequencing to allow differentiation between paralogs arising from the recent salmonid genome duplication. 5′ (reverse) sequencing was attempted on 7487 of the 71,144 *S. salar* clones. The 80,388 high-quality ESTs (55,082 forward and 6737 reverse *S. salar*, 14,544 forward *O. mykiss*, 1317 forward *O. tshawytscha*, 1243 forward *O. nerka*, and 1465 forward *C. clupeaformis*) were assembled by using PHRAP under high stringency to identify EST clusters (contiguous sequences, or contigs) representing redundant transcripts (Tables 1, 2). The average trimmed PHRED20 length of these ESTs is 546 bases. The 61,819 *S. salar* ESTs were assembled into 11,560 contigs (with 17,150 singletons remaining), 14,544 *O. mykiss* ESTs formed 2370 contigs (6611 singletons), 1317 *O. tshawytscha* ESTs formed 136 contigs (949 singletons), 1243 *O. nerka* ESTs formed 291 contigs (229 singletons), and 1465 *C. clupeaformis* ESTs formed 138 contigs (1038 singletons; Table 1). There are 28,710 assembled *S. salar* sequences (putative transcripts), 8981 *O. mykiss* putative transcripts, 1085 *O. tshawytscha* putative transcripts, 520 *O. nerka* putative transcripts, and 1176 *C. clupeaformis* putative transcripts (Table 1). Results of alternate assemblies (CAP3 and stackPACK) of this EST collection are available at http://web.uvic.ca/cbr/grasp. The largest *S. salar* contig contains 252 ESTs (prolactin); the largest *O. mykiss* contig is size 93 (parvalbumen β); the largest *O. tshawytscha* contig is size 10 (cytochrome c oxidase subunit II); the largest *O. nerka* contig is size 21 (similar to ribosomal protein L41); and the largest *C. clupeaformis* contig is size 28 (ependymin; Table 1). BLAST alignments of ESTs against combined ribosomal and mitochondrial sequence databases (see Methods) identified 1052 *S. salar*, 396 *O. mykiss*, 103 *O. tshawytscha*, 40 *O. nerka*, and 157 *C. clupeaformis* reads.

Preliminary analysis of aligned *S. salar* and *O. mykiss* assembled ESTs identifies 1892 sequence pairs with >80% identity (see Methods). Of these, 1429 (~76%) were contained within a distinct peak from 90%–97% identity (average ~94%) at the nucleotide level. As it is difficult to distinguish orthologs from sequence pairs related by paralogy resulting from gene or genome duplications, a more focused study is underway.

REPuter (Kurtz et. al. 2001) identifies 11.9% of the total length of assembled sequences (TLAS) as known classes of repeats; 6.7% of the TLAS is composed of SINEs (predominately HpaI), whereas satellites, pseudogenes (including a large number of transposable element-associated sequences), and transposable elements account for 3.4%, 1.1%, and 0.7% of the TLAS, respectively.

Library Complexity and Gene Discovery

By using the March 3, 2003, versions of our EST database and GenBank databases, each library's ESTs were BLASTN- and BLASTX-aligned against a database composed of all nonredundant nucleic and amino acid sequences from that species in GenBank plus our collection of nonredundant ESTs. Percentage of singleton values for each library were calculated by using the August 25, 2003, version of the GRASP database (Table 2). Higher "percent new," higher "percent no significant BLAST hit," and higher "percent singleton" values indicate higher rates of gene discovery and higher complexity in a given library. For several of our libraries, higher- and lower-molecular-weight (MW) fractions were cloned separately. By using all three metrics, higher-MW fractions are of higher complexity (and higher rates of new gene identification) than their corresponding lower-MW fractions (Table 2). For example, the lower-MW *S. salar* head kidney library (average insert size of 1031 bp) has values of 15.0% new, 12.7% no BLAST hit, and 12.0% singletons, whereas the corresponding higher-MW library (average insert size of 2307 bp) values are 38.2%, 38.4%, and 35.3% respectively (Table 2). In addition, different suites of genes are identified in lower- and higher-MW fractions of a single cDNA library. This qualitative difference is evident in a list of the largest EST clusters in select non-normalized libraries/library groups (Table 3). Excluding ribosomal and mitochondrial clusters, the most abundant transcripts in the *S. salar* pyloric caecum lower-MW library are apolipoprotein A-I (2 forms of the gene in separate EST clusters), apolipoprotein E, 28 kD − 1e apolipoprotein, and galectin, whereas the largest EST contigs in the associated higher-MW library are selenoprotein Pa, MHC class I heavy chain, meprin A α, an unknown, and type II keratin E2, (Table 3). Likewise, in the head kidney, different sets of highly prevalent transcripts are seen in lower- and higher-MW library groups (Table 3). These results indicate that the preparation and characterization of higher-MW fractions of cDNA libraries improved the rate of gene discovery in the GRASP EST project.

Insert orientation in various types of cDNA library was analyzed to determine its potential influence on gene discovery

Table 1. Salmonid EST Project Summary Statistics[a]

	Atlantic salmon[b]	Rainbow trout[c]	Chinook salmon[d]	Sockeye salmon[e]	Lake whitefish[f]
Number of good sequences[g]	61,819[h]	14,544	1317	1243	1465
Average trimmed EST length (bp)[i]	563	484	492	456	486
Number of contigs[j]	11,560	2370	136	291	138
Number of singletons	17,150	6611	949	229	1038
Number of putative transcripts	28,710	8981	1085	520	1176
Max. assembled sequence size (no. of ESTs)	252	93	10	21	28
Average assembled sequence size (no. of ESTs)	2.15	1.61	1.21	2.39	1.24
Number of assembled ESTs with[k]					
Significant BLASTX hits	10,511	3562	239	337	253
Significant BLASTN hits	13,459	4337	462	331	466
No significant BLAST hits	11,802	3667	566	118	663
Percentage with no significant BLAST hits[k]	41.1	40.8	52.2	22.7	56.4
Number of contigs containing[i]					
2 ESTs	5606	1360	90	108	97
3 ESTs	2322	454	26	96	21
4–5 ESTs	2030	350	12	48	9
6–10 ESTs	1149	145	8	32	8
11–20 ESTs	331	41	0	6	1
21–30 ESTs	67	12	0	1	2
31–50 ESTs	36	4	0	0	0
>50 ESTs	19	4	0	0	0

[a]Assembled from the March 3, 2003, version of the GRASP EST database using PHRAP. Results of CAP3 and stackPACK assemblies of the March 3, 2003 GRASP EST database are available at http://web.uvic.ca/cbr/grasp
[b]*Salmo salar*
[c]*Oncorhynchus mykiss*
[d]*Oncorhynchus tshawytscha*
[e]*Oncorhynchus nerka*
[f]*Coregonus clupeaformis*
[g]A sequence is considered "good" if its trimmed PHRED20 length is at least 100 bases.
[h]Includes 55.082 good forward (3′) and 6737 good reverse (5′) reads. Of 5606 good reverse reads from clones with good forward reads, 2268 overlap/cluster with the corresponding forward reads.
[i]Vector, low-quality, and contaminating bacterial sequences are trimmed.
[j]A contig (contiguous sequence) contains two or more ESTs.
[k]Threshold for BLASTN and BLASTX significance: 10^{-5}

rates. All libraries in this database were classified by type (i.e. normalized, subtracted), and insert orientations in two libraries from each class were determined (see Supplemental table at http://web.uvic.ca/cbr/grasp for data, method, and discussion of bias). Incidences of reverse-oriented inserts were as follows: 4.5% in non-normalized, nonfractionated libraries (average of two analyzed libraries' values); 29% in non-normalized, higher-MW libraries; 20% in non-normalized, lower-MW libraries; 10% in normalized libraries; and 71.5% in subtracted (randomly cloned) libraries. The weighted average across all four directionally cloned library types (contributing 84.9% of the ESTs in the database) is 9.1% reverse orientation. M13 forward-read ESTs from reverse-oriented inserts give 5′ sequence.

The somewhat higher incidence of reverse-oriented inserts in higher-MW fraction libraries might contribute to the higher "% new" and "% singleton" values of these libraries over their lower-MW counterparts in our database (Table 2). However, insert orientation differences between library classes do not explain the dramatically higher "% no BLAST hit" values seen in higher-MW libraries (Table 2). Because most EST projects contributing to GenBank databases are biased toward 5′ sequencing, the higher "% no BLAST hit" values of our higher-MW libraries are likely conservative indices of the elevated rates of gene discovery associated with these libraries.

Assembled *S. salar* and *O. mykiss* ESTs were checked for open reading frames (ORFs) >200 bp (Fig. 1A). The chance of a random 66 codon (198 bp) ORF is $(61/64)^{66} = 0.04206$ ($P < 0.05$). Most of our ESTs are 3′ reads. The average observed 3′ UTR in this database is 264 bases (60 3′ ESTs considered; range, 59 to 592 bases), and average trimmed EST lengths are 484 to 563 bases (Table 1). Therefore, we believe that screening for 200-bp ORFs allows for adequate evaluation of the coding portion of the ESTs without excessive bias against genes with longer 3′ UTRs.

Of the 28,710 assembled *S. salar* sequences, 22,622 (79%) have ORFs >200 bp (Fig. 1A). Of these, 10,123 (45%) have significant ($E < 10^{-5}$) BLASTX hits, and 9822 (43%) have significant ($E < 10^{-5}$) BLASTN hits (Fig. 1A). Novel salmonid genes may be included in the 12,499 assembled ESTs containing 200-bp ORFs but without BLASTX matches (Fig. 1A). Of the 6088 assembled *S. salar* ESTs without 200-bp ORFs, 388 (6%) have significant BLASTX hits (likely representing cDNAs coding for short proteins) and 1664 (27%) have significant BLASTN hits (likely representing cDNAs for short proteins as well as previously identified salmonid intronic and untranslated sequences; Fig. 1A). The 4424 assembled *S. salar* ESTs having neither 200-bp ORFs nor BLASTN hits (Fig. 1A) probably include novel salmonid cDNAs with long 3′ UTRs. The ORF and BLAST results for *O. mykiss* assembled sequences are very similar to those for *S. salar* (Fig. 1B).

Using Functional Annotation to Infer Putative Organ Functions

The gene ontology (GO) statistics presented in this report reflect the state of our database on March 3, 2003. For the collective *S. salar* libraries, and for each *S. salar* library group, assembled ESTs were assigned putative molecular functions based on BLASTX similarity to functionally annotated human protein sequences (Gene Ontology Consortium 2001).

To illustrate ways in which this salmonid EST database may be mined for information on putative organ functions, we focused on a selection of *S. salar* organ-specific library groups: gill, mixed gut (stomach + mid-gut + hind-gut, not including pyloric caecum), ovary, pyloric caecum, and pituitary gland (Table 4). Overall, 26% of *S. salar* assembled ESTs matched sequences in the GO database (Table 4). For organ-specific libraries or library groups, the percentage of assembled ESTs hitting GO sequences ranged from 25% (ovary) to 40% (pyloric caecum; Table 4). Z-statistics were used to determine if, for a given GO classification, the proportion of assembled ESTs in an organ-specific *S. salar* library group differed significantly from the proportion of assembled ESTs from remaining *S. salar* library groups (see Methods, Table 4, and Supplemental data at http://web.uvic.ca/cbr/

Table 2. Salmonid cDNA Library Summary Statistics[a]

Library/library group	Average insert	No. good seq.[b]	No. putative trans.[c]	Max. contig.	Ave. contig	% new (spec)[d]	% no BLAST[e]	% single[f]	No. on chip
Salmo salar									
Brain	1413 bp	1161	891	23	1.30	38.4	47.3	30.4	0
Normalized heart	494 bp	729	634	5	1.14	30.4	29.8	29.2	0
Esophagus[g]	1908 bp	749	506	51	1.48	31.2	32.2	20.4	2
Gill	840 bp	2308	1751	14	1.31	30.6	35.7	24.9	0
Head kidney lower MW	1031 bp	784	425	68	1.84	15.0	12.7	12.0	0
Head kidney higher MW	2307 bp	867	750	14	1.15	38.2	38.4	35.3	6
Head kidney, infected[h]	895 bp	1921	1173	24	1.63	65.8	36.1	35.0	315
Head kidney, stimulated[i]	984 bp	1233	898	71	1.37	50.5	32.3	39.8	0
Normalized liver	619 bp	1379	977	16	1.41	24.5	22.6	21.2	362
Mixed gut[j]	1450 bp	3753	2509	41	1.49	29.8	29.9	20.1	1
Norm, skeletal muscle	928 bp	903	770	5	1.17	32.5	30.9	29.5	0
Ovary	723 bp	2664	2239	17	1.18	40.2	45.7	33.0	0
Pituitary gland	692 bp	2883	1512	123	1.90	34.7	35.8	17.7	214
Pyloric caecum lower MW	1043 bp	329	251	16	1.31	11.9	15.5	10.6	2
Pyloric caecum higher MW	2400 bp	716	564	16	1.26	39.8	36.2	30.4	24
Norm pyl.caecum low MW[k]	884 bp	5514	3082	37	1.78	36.3	25.5	14.8	248
Retina[l]	662 bp	1118	914	17	1.22	33.1	41.0	29.1	0
Normalized mixed tissue[m]	1556 bp	24,534	15,458	29	1.58	65.3	39.7	34.7	1361
Spleen lower MW	1278 bp	210	173	5	1.21	17.9	17.3	20.5	0
Spleen higher MW	2089 bp	1926	1575	22	1.22	42.4	41.8	35.9	79
Testes[n]	664 bp	2038	1530	14	1.33	34.2	41.4	24.5	2
Whole juvenile	1046 bp	268	225	6	1.19	21.3	22.2	21.6	3
Normalized whole juvenile	883 bp	3832	2494	47	1.53	37.4	31.1	20.5	500
Oncorhynchus mykiss									
Brain	805 bp	330	284	8	1.16	61.6	39.1	50.9	0
Mixed embryonic[o]	779 bp	1815	1314	34	1.38	78.2	47.9	50.8	0
Mixed gonad[p]	574 bp	6956	4862	58	1.43	83.4	40.0	52.5	279
Whole juvenile lower MW	1087 bp	296	234	6	1.26	37.1	15.4	31.8	65
Whole juvenile higher MW	2181 bp	936	759	17	1.23	70.0	41.6	55.0	71
Norm. whole juv. lower MW	962 bp	4211	2335	91	1.80	71.6	32.2	28.8	23
Oncorhynchus tshawytscha									
Mixed tissue lower MW	1025 bp	576	498	8	1.15	83.5	46.2	60.1	0
Mixed tissue higher MW	1294 bp	741	660	5	1.12	87.8	52.3	64.8	0
Coregonus clupeaformis									
Brain lower MW	802 bp	729	583	16	1.25	87.6	52.0	53.9	0
Brain higher MW	1203 bp	736	649	18	1.13	90.2	58.4	67.0	0

Shading indicates statistics related to complexity and gene discovery rate in higher- and lower-molecular weight (MW) cDNA libraries discussed in text.

[a]Compiled using the March 3, 2003, version of the GRASP EST database (except percent singletons[f])

[b]Sequence considered "good" if its trimmed PHRED20 length is at least 100 bases.

[c]Number of putative transcripts (assembled ESTs) = number of contigs + number of singletons.

[d]Percent new (species) values, used to estimate gene discovery rate, is the number of previously unidentified EST clusters (including contigs and singletons) divided by the total number of clusters.

[e]Percentage of EST clusters (including contigs and singletons) with no significant BLASTN or BLASTX hit ($E < 10^{-5}$).

[f]Percent singletons, calculated using the August 25, 2003, version of the GRASP EST database, was the number of singletons in a library (considering all data in the database) divided by the total number of ESTs ("good sequences") from that library.

[g]Esophagus library group contains three size fractions of a single cDNA library.

[h]Infected head kidney library group contains four size fractions of a single suppression subtractive hybridization (SSH) library. Includes 82 good reverse reads, 76 with good paired (from the same clone) forward reads (18 overlapping/clustering together).

[i]Stimulated head kidney library group contains four size fractions of a single SSH library and two size fractions of a single cDNA library.

[j]Mixed gut (stomach, mid-gut, hind-gut) library group contains six separately cloned size fractions of a single cDNA library.

[k]Includes 1949 good reverse (5′) reads, 1804 of which have good paired forward reads (1205 overlapping).

[l]Includes 143 good reverse (5′) reads without attempted forward reads.

[m]Normalized mixed tissue (spleen, kidney, brain) includes 3929 good reverse (5′) reads, 3503 with good paired forward reads (871 overlapping).

[n]Includes 318 good reverse (5′) reads without attempted forward reads. Of 316 good reverse reads with attempted forward reads, 223 have good paired forward reads (174 overlapping). This library was contaminated with ovary and retina cDNAs during gel fractionation.

[o]Mixed embryonic library group contains four SSH libraries and two cDNA libraries from different stages.

[p]Mixed gonad library group contains 12 SSH libraries and 8 cDNA libraries.

Table 3. Largest EST Clusters in Select Single-Tissue, Nonnormalized Atlantic Salmon Libraries/Library Groups[a]

Tissue (library)[a]	Total ESTs	ESTs in cluster	BLAST[b] E-value	Length (% identity)[c]	GenBank hit acc. no.[b]	Gene identification (species) of top BLAST hit[b]
Brain	1161	23	(X) 1.5E-60	113 (95.6%)	P28770	Ependymin I (*Oncorhynchus mykiss*)
		19	(X) 2.9E-128	221 (99.5%)	P28772	Ependymin II (*Salmo salar*)
		17	(X) 7.2E-10	119 (29.1%)	O62680	Membrane attack complex inhibition factor (*Sus scrofa*)
		15	(X) 5.6E-6	97 (32.9%)	P06906	Myelin basic protein (*Pan troglodytes*)
		8	n/a[d]	n/a[d]	n/a[d]	unknown
Esophagus	749	88	(X) 1.3E-24	111 (58.9%)	BAA86981	Novel member of chitinase family (*Homo sapiens*)[e]
		32	(X) 4.8E-59	315 (59.9%)	BAA86981	Novel member of chitinase family (*H. sapiens*)[e]
		9	(X) 4.7E-58	154 (75.3%)	AAD56283	Pepsinogen A form IIa (*Pseudopleuronectes americanus*)
		8	(X) 0	461 (99.5%)	AAG38613	Elongation factor 1 α (*S. salar*)
		8	(X) 9.3E-83	175 (97.7%)	CAA37852	Creatine kinase (*O. mykiss*)
Gill	2308	14	(X) 1.6E-64	116 (98.2%)	AAG17525	β-2 microglobulin (*S. salar*)
		13	(X) 5.6E-132	244 (100%)	CAA49726	MHC Class II β chain (*S. salar*)
		11	(X) 6.6E-22	62 (62.9%)	AAG30024	C-type lectin 2-1 (*O. mykiss*)
		10	(N) 0	346 (99.1%)	BG936357	unknown spleen clone SS1-0729 (*S. salar*)
		9	(N) 0	876 (98.4%)	BG934637	unknown kidney clone SK1-0954 (*S. salar*)
Head kidney (lower MW)	784	57	(X) 1.2E-79	148 (94.5%)	CAA65945	β globin^{A} (*S. salar*)[f]
		53	(X) 1.1E-83	148 (100%)	CAA49580	β globin^{B} (*S. salar*)[f]
		41	(X) 2.0E-82	148 (97.9%)	CAA65945	β globin^{C} (*S. salar*)[f]
		7	(X) 1.1E-83	148 (100%)	CAA65945	β globin^{D} (*S. salar*)[f]
		7	(X) 4.6E-66	119 (100%)	CAA65944	α globin (*S. salar*)
Head kidney (higher MW)	867	24	(X) 9.4E-146	252 (100%)	O42161	β actin (*S. salar*)
		5	(X) 0	443 (96.2%)	A46533	Immunoglobulin heavy chain constant region (*S. salar*)[g]
		4	(X) 1.1E-141	287 (100%)	S21175	dnaK-type molecular chaperone hsc71 (*O. mykiss*)
		4	(X) 0	443 (100%)	A46533	Immunoglobulin heavy chain constant region (*S. salar*)[g]
		4	(N) 0	695 (99.7%)	AJ424426	unknown kidney clone k09F03 (*S. salar*)
Mixed gut (stomach, mid, & hind)	3753	47	(X) 1.9E-85	239 (79.1%)	JH0472	Apolipoprotein A-I (*S. salar*)
		20	(X) 0	330 (97.8%)	AAK69705	Procathepsin B (*O. mykiss*)
		18	(X) 3.8E-34	76 (92.1%)	CAC45057	Type II keratin E2 (*O. mykiss*)
		17	(X) 0	461 (99.5%)	AAG38613	Elongation factor 1 α (*S. salar*)
		16	(X) 4.8E-89	325 (51.6%)	CAC87888	Toad pancreatic chitinase (*Bufo japonicus*)
Ovary	2664	13	(X) 1.2E-48	139 (59.7%)	AAO43606	Serum lectin isoform 2 (*S. salar*)[h]
		11	(X) 4.3E-48	139 (59.1%)	AAO43606	Serum lectin isoform 2 (*S. salar*)[h]
		6	n/a[d]	n/a[d]	n/a[d]	unknown
		5	n.a[d]	n/a[d]	n/a[d]	unknown
		5	(X) 1.5E-16	127 (41.0%)	P56733	Avidin-related protein 3 (*Gallus gallus*)
Pituitary Gland	2883	123	(N) 0	545 (98.5%)	X84787	Prolactin (*S. salar*)
		96	(N) 6.5E-166	1084 (94.9%)	X69809	Proopiomelanocortine B (*O. mykiss*)
		48	(N) 5.9E-73	173 (94.7%)	X69808	Proopiomelanocortine A (*O. mykiss*)
		44	(X) 1.0E-26	50 (100%)	AAA49558	Growth hormone (*S. salar*)
		43	(X) 3.1E-60	114 (100%)	AAA49407	Gonadotropin-I α (*Oncorhynchus keta*)
Pyloric caeca (lower MW)	329	8	(X) 5.1E-123	262 (98.0%)	AAA88542	Apolipoprotein A-I (*S. salar*)[i]
		7	(X) 9.9E-67	268 (49.2%)	CAB65320	Apolipoprotein E (*O. mykiss*)
		5	(X) 9.7E-55	254 (46.0%)	BAB40965	28 kD-1e apolipoprotein (*Anguilla japonica*)
		5	(X) 8.1E-27	129 (44.1%)	AAF61069	Galectin (*Paralichthys olivaceus*)
		4	(X) 3.9E-100	238 (92.4%)	JH0472	Apolipoprotein A-I (*S. salar*)[i]
Pyloric caeca (higher MW)	716	16	(X) 1.2E-48	168 (58.2%)	AAG53688	Selenoprotein Pa (*Danio rerio*)
		10	(X) 4.5E-141	286 (86.4%)	AAG02508	MHC Class I heavy chain (*O. mykiss*)
		8	(X) 1.4E-32	122 (53.2%)	AAL85339	Meprin A α; PABA peptide hydrolase (*H. sapiens*)
		7	(N) 2.0E-12	115 (84.3%)	BG935597	unknown liver clone SL1-0959 (*S. salar*)
		6	(X) 3.8E-34	76 (92.1%)	CAC45057	Type II keratin E2 (*O. mykiss*)

[a]Compiled using the August 16, 2003 version of the GRASP EST database and excluding ribosomal and mitochondrial EST clusters. For notes on libraries and library groups, see Table 2.
[b]Most significant BLASTN (N) or BLASTX (X) hit is reported. BLASTX hit reported if top BLASTN hit not associated with a named gene.
[c]Extent of BLAST hit aligned region, and percent identity over the aligned region. Length (and percent identity) refers to amino acids if BLASTX reported, and nucleic acids if BLASTN reported.
[d]Not applicable, as there are no significantly similar ($E < 10^{-5}$) sequences in non-redundant GenBank nucleotide or amino acid sequence databases.
[e]The C-terminal 111 amino acids of the aligned translations of these 2 EST contigs are 85.6% identical, indicating 2 distinct forms of the gene.
[f]The aligned translations of β globins A, B, and C differ from the translation of β globin D at 8 of 148 (94.6% identity), 2 of 148 (98.6% identity), and 3 of 148 (98.0% identity) residues, respectively.
[g]The aligned translations of these two EST clusters differ at 17 of 443 residues (96.2% identity).
[h]The aligned translations of these two EST clusters differ at four of 139 residues (97.1% identity). They are ~60% identical, at the amino acid level, to a third ovary lectin cluster containing four ESTs (not shown in table).
[i]The aligned translations of these two EST clusters (238 amino acids) are 84.0% identical.

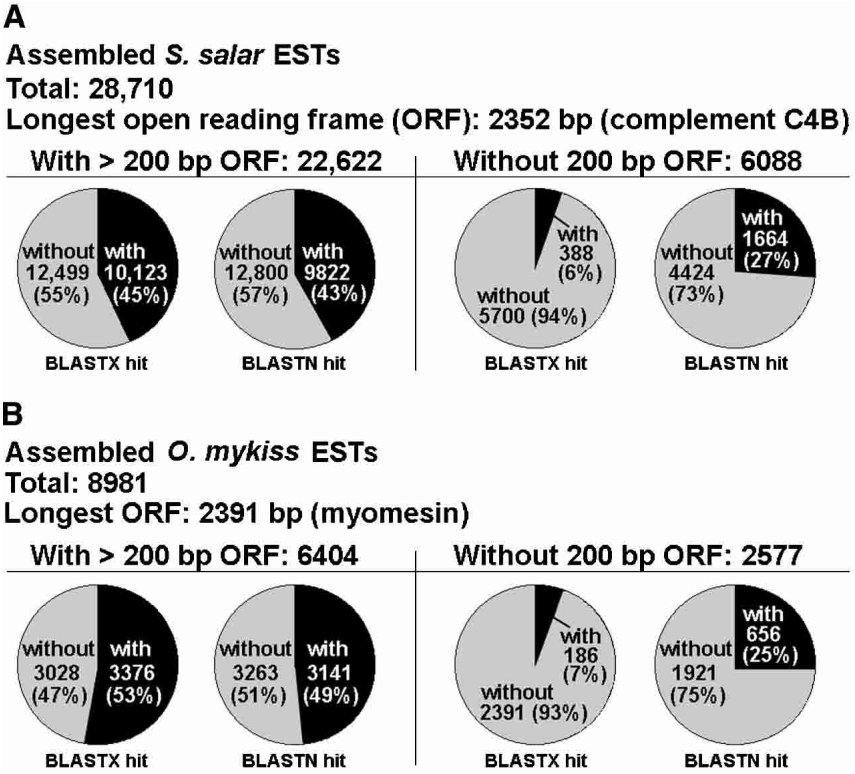


Figure 1 Open reading frame (ORF) and BLAST results. Numbers of assembled *Salmo salar* (*A*) and *Oncorhynchus mykiss* (*B*) ESTs with and without 200 base ORFs are given. Within each of these categories, proportions of assembled ESTs with and without significant ($E < 10^{-5}$) BLASTX hits against GenBank nonredundant protein database are shown, as are proportions of assembled ESTs with and without significant ($E < 10^{-5}$) BLASTN hits against the nonredundant nucleotide database. The lengths and putative identifications (gene names of best BLASTX hits) of the longest ORFs in each species are given.

grasp). Because putative organ functions are sought, only those GO categories with disproportionately high numbers of assembled ESTs will be discussed. The gill library has disproportionately high numbers of assembled ESTs in the "iron binding," "oxidoreductase, acting on heme group of donors," and "transporter" GO categories (Table 4). Disproportionately high numbers of ovary assembled ESTs are seen in GO categories related to heavy metal (copper, iron, and zinc) binding and enzyme inhibition (Table 4). The pituitary gland library has disproportionately high numbers of assembled ESTs in "iron binding," "hormone binding," and "oxidoreductase, acting on heme group of donors" categories (Table 4).

This approach was used to acquire putative functional information on a poorly characterized organ, the pyloric caecum. The teleost pyloric caecum, a large elaborate set of finger-like extensions off the gut, is known to play an important role in nutrient uptake (Buddington and Diamond 1986). There is scant literature on the fish pyloric caecum. Douglas et al. (1999b) classified 147 winter flounder pyloric caecum ESTs by putative function. Winter flounder pyloric caecum libraries were used to isolate cDNA clones encoding trypsinogen (Douglas and Gallant 1998) and aminopeptidase N (Douglas et al. 1999a). Here we report 6559 *S. salar* pyloric caecum ESTs, facilitating a deeper understanding of gene expression in this organ.

Both mixed gut and pyloric caecum library groups have disproportionately high numbers of assembled ESTs in the "enzyme," "oxidoreducatase, acting on heme group of donors," and "transporter" GO categories (Table 4). These may point to general functions along the digestive tract. "Iron binding" and "hydrolase" GO categories have disproportionately high numbers of assembled ESTs in mixed gut but not pyloric caecum (Table 4). There are disproportionately high numbers of assembled ESTs in the "cytochrome P450," "selenium binding," "oxidoreductase, acting on NADH or NADPH," "oxidoreductase, acting on peroxide as acceptor," and "transferring sulfur-containing groups" categories in pyloric caecum but not mixed gut library groups (Table 4), indicating putative specialized roles for the pyloric caecum. Selenium is a component of selenoprotein P and glutathione peroxidases, antioxidant enzymes that protect cells from oxidative injury (Deplancke and Gaskins 2002; Burk et al. 2003; Schomburg et al. 2003). Selenoprotein P is one of the largest EST contigs in the pyloric caecum library group (Table 5), and this library group contains at least eight different assembled ESTs identified by BLAST as glutathione peroxidases. At least 10 different pyloric caecum assembled ESTs are identified as cytochromes P450, a class of heme-containing monooxygenases involved in metabolism of foreign compounds such as environmental pollutants and agrochemicals (Danielson 2002). Collectively, these results indicate that the salmon pyloric caecum functions in redox control and as a barrier against the systemic uptake of xenobiotics.

Additional hypothetical functions of the pyloric caecum may be proposed by examining the largest EST clusters (representing highly expressed genes) in the pyloric caecum library group, and locating other members of these clusters across all *S. salar* library groups (Table 5). Several defense-relevant EST clusters, including CC chemokine macrophage inflammatory protein (MIP)-3a, galectin, and GDP-D-mannose-4,6-dehydratase (GMD), derive most of their ESTs from pyloric caecum libraries (Table 5). Galectins serve as master regulators of immune cell homeostasis during innate immune responses (Rabinovich et al. 2002). GMD is required for the synthesis of fucosylated oligosaccharides, selectin ligands involved in leukocyte extravasation (Ohyama et al. 1998; Eshel et al. 2001). These data indicate an innate defense function of the salmonid pyloric caecum. That previously unknown EST types, frequencies, and distributions have been observed among pyloric caecum and other organs by this analysis indicates general utility for this approach in revealing unknown functions of many other organ and cellular systems.

Application of a Salmonid cDNA Microarray to Different Species

A preliminary cDNA microarray (available from corresponding author), composed of 6440 AS and 916 rainbow trout (RT) cDNA elements or spots (Table 6), was hybridized with labeled targets from three members of the order Salmoniformes (AS, RT, and lake whitefish [LW]) and one member of the order Osmeriformes (rainbow smelt; Fig. 2A) to explore the validity of using this microarray with other fish species. Hybridization performance of each species' labeled target to the salmonid elements was judged from the numbers of AS and RT elements passing a hybridization signal threshold, and mean total raw signals from AS and RT elements (see Methods; Table 6). No transformations or normal-

Table 4. Gene Ontology[a] (Molecular Function) of Assembled ESTs From Organ-Specific Libraries/Library Groups[b]

Library/library group:[c]	All[d] (26%)	Gill[e] (31%)	MG[f] (35%)	Ovary[g] (25%)	PC[h] (40%)	Pituit.[i] (34%)
GO term[a]						
All molecular function	6937	444	722	430	1192	441
Antioxidant	10	2	0	1	2	0
Apoptosis regulator	26	2	1	0	4	0
Binding	3328	***275***	331	***245***	***525***	***309***
Heavy metal binding	194	19	28	***42***	25	18
Copper binding	*7*	0	0	***2***	1	0
Iron binding	55	***8***	***17***	***24***	8	***9***
Zinc binding	129	11	11	***16***	16	9
Oxygen binding	32	3	2	0	9	0
Cytochrome P450	21	0	2	0	***8***	0
Protein binding	*724*	51	***39***	47	*101*	***17***
Transcription factor binding	166	*7*	***3***	6	24	***2***
Receptor binding	177	8	14	*7*	31	***37***
Cytokine	40	5	2	0	8	1
Hormone	52	1	4	1	9	***30***
Selenium binding	9	0	3	0	***6***	0
Cell adhesion molecule	44	1	8	1	4	2
Chaperone	148	6	21	10	23	11
Heat shock protein	53	3	9	*6*	9	6
Defense/immunity protein	192	*7*	13	*4*	***13***	***0***
Antiviral response protein	29	1	3	1	5	0
Enzyme	2484	***110***	***285***	***110***	***486***	***80***
Hydrolase	1116	***37***	***154***	***39***	212	***26***
Kinase	292	***6***	***12***	***7***	*34*	***4***
Oxidoreductase	478	***44***	***66***	38	***127***	37
Disulfide oxidoreductase	20	1	0	1	5	2
Oxidoreductase, acting on heme group of donors	66	***15***	***16***	8	***19***	***11***
Oxidoreductase, acting on CH-OH group of donors	89	4	10	*0*	22	2
Oxidoreductase, acting on NADH or NADPH	125	11	15	12	***32***	12
Oxidoreductase, acting on peroxide as acceptor	*26*	2	6	3	***11***	2
Transferase	595	***15***	***38***	***21***	98	***9***
Transferring phosphorus-containing groups	338	***8***	***15***	*12*	44	***6***
Transferring sulfur-containing groups	16	0	1	0	***6***	0
Enzyme regulator	288	17	22	25	56	*7*
Enzyme activator	123	*7*	11	11	28	*1*
Enzyme inhibitor	167	11	12	***19***	28	5
Transcription regulator	517	***14***	***15***	22	*66*	***11***
Transcription factor	427	***9***	***13***	18	***49***	***10***
Transcription cofactor	145	6	***3***	6	22	***1***
Translation regulator	150	8	23	11	30	*6*
Transporter	754	***62***	***109***	47	***198***	49

Z-statistics and associated *P* values were used to compare sample proportions (see Methods). Values with negative Z-statistics have fewer assembled ESTs than expected (italics = $P < 0.05$, bold italics = $P < 0.01$). Values with positive Z-statistics have more assembled ESTs than expected (10% grey = $P < 0.05$; 25% grey = $P < 0.01$).

[a]Classified using guidelines of the Gene Ontology Consortium 2001 (http://www.geneontology.org). Indented terms are children of the above parent term. A selection of GO categories is presented.

[b]For notes on libraries and library groups, see Table 2. ESTs were assembled by using PHRAP. GO statistics were compiled using the March 3, 2003, version of the GRASP EST database. Percentages are for assembled ESTs hitting GO sequences.

[c]GO hits in all three categories (Molecular Function, Cellular Component, and Biological Process) are pooled for this calculation.

[d]All *S. salar* includes 61,819 ESTs; 6937 assembled ESTs having significant BLASTX hits ($E < 10^{-5}$) are annotated by molecular function. "All *S. salar*" are inflated due to 1131 reverse (5′) reads without good corresponding forward reads, and 3338 reverse reads that do not overlap/cluster with corresponding good forward reads.

[e]Of 2308 Gill library ESTs 444 assembled ESTs with significant BLASTX hits are annotated by molecular function.

[f]Of 3753 Mixed Gut library ESTs, 722 assembled ESTs with significant BLASTX hits are annotated by molecular function.

[g]Of 2664 Ovary library ESTs, 430 assembled ESTs with significant BLASTX hits are annotated by molecular function.

[h]PC = all pyloric caecum libraries (lower MW, higher MW, and normalized). Of 6559 pyloric caecum library ESTs, 1192 assembled ESTs with significant BLASTX hits are annotated by molecular function. PC values are slightly inflated due to 145 reverse reads without good corresponding forward reads, and 599 reverse reads that do not overlap/cluster with good corresponding forward reads.

[i]Of 2883 pituitary gland library ESTs, 441 assembled ESTs with significant BLASTX hits are annotated by molecular function.

izations were performed on the data. Data and statistics for all slides are given in Table 6.

To evaluate the effect of element (cDNA spotted onto the microarray slide) and target (labeled cDNA hybridized to the slide) species affiliations on hybridization characteristics, data and statistics for AS and RT microarray elements were compiled separately. On AS probes, AS target gave the highest signal (mean of three slides: 2.01E7, SEM 4.99E5), followed by RT (mean of three slides: 1.88E7, SEM 8.16E4), LW (mean of three slides: 1.54E7, SEM 3.31E5), and rainbow smelt (mean of three slides: 6.61E6, SEM 5.37E5; Table 6). On RT probes, RT target gave the highest signal (mean of three slides: 2.53E6, SEM 4.19E4), fol-

Table 5. Largest EST Clusters[a] in the Pyloric Caecum Library Group,[b] and Locations of Other Members of Clusters Across *S. salar* cDNA Library Groups[b]

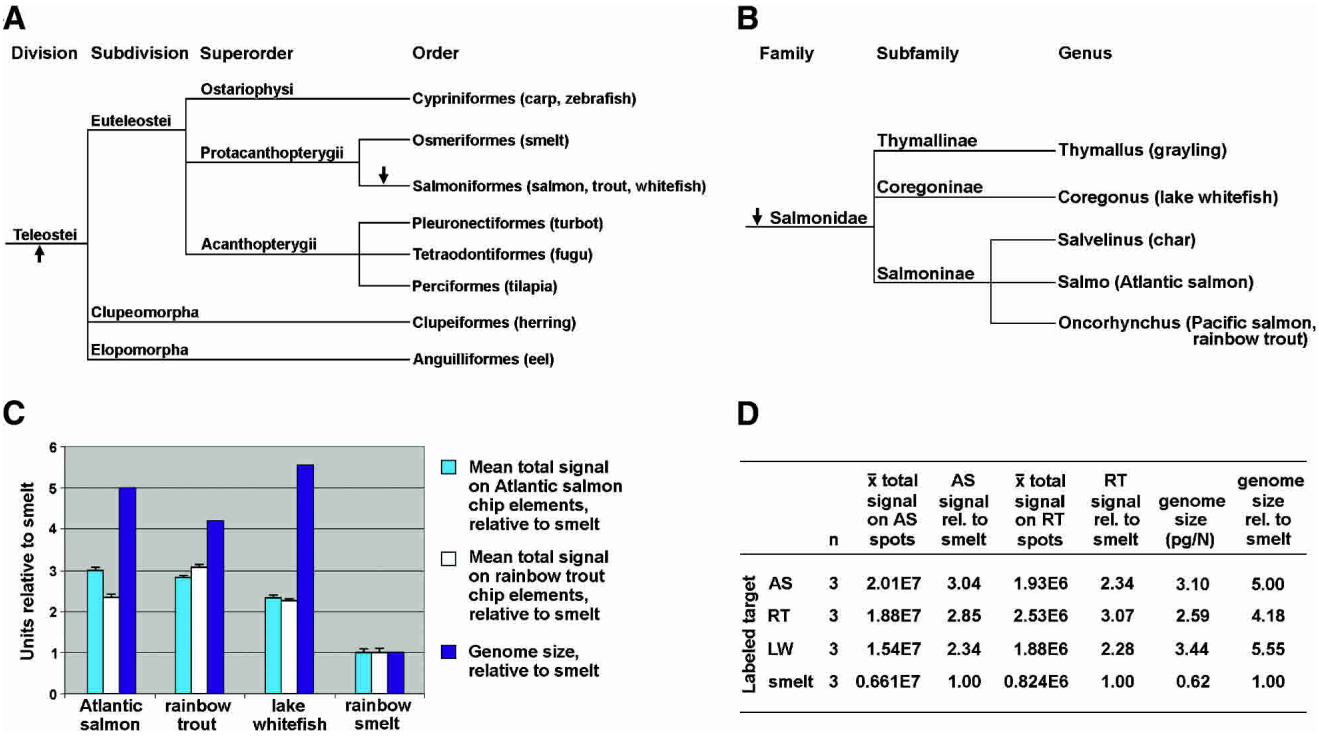
GenBank hit acc. no.[c]	Gene identification (species) of top BLAST hit[c]	BLAST[c] E-value	Length (% identity)[c]	pc	br	es	gi	he	ki	li	mg	ov	pi	re	skb	sp	te	w
CAA49679	Trypsin III (*Salmo salar*)	(X) 5.0E-130	230 (100%)	40_{14}							1				4_1			3
CAB65320	Apolipoprotein E (*Oncorhynchus mykiss*)	(X) 9.9E-67	268 (49.2%)	34_9							11							1
P80961	Antifreeze protein LS-12 (*M. octodecemspinosus*)	(X) 5.4E-11	126 (41.2%)	27_4	1					1	7				2			4
AAG53688	Selenoprotein Pa (*Danio rerio*)	(X) 1.1E-48	167 (58.2%)	21_2					2		8			1	6			3
I51348	MHC class I (*S. salar*)	(X) 4.4E-103	230 (70.5%)	16_2			6	2	1		1					17		
BAB40965	28 kDa-1e apolipoprotein[A] (*Anguilla japonica*)[d]	(X) 2.7E-39	234 (38.4%)	16_4							3							
AAC52505	CRP ductin α (*Mus musculus*)	(X) 1.1E-81	412 (41.4%)	14_3							1							
CAA10948	CC Chemokine MIP-3a (*M. musculus*)	(X) 7.9E-8	79 (37.8%)	14_5							1							
CAB46819	Splicing factor (*Canis familiaris*)	(X) 1.7E-40	84 (92.8%)	13_5											3		1	4
AAL85339	Meprin A α (*Homo sapiens*)	(X) 8.0E-107	340 (53.0%)	12_2							4							
BAA82366	Chymotrypsinogen 2 (*Paralichthys olivaceus*)	(X) 5.5E-115	254 (81.1%)	12_4							1							3
BAA82370	Elastase 4 precursor (*P. olivaceus*)	(X) 0	261 (85.4%)	12_4							1							2
CAC45057	Type II keratin E2 (*O. mykiss*)	(X) 3.8E-34	75 (92.1%)	11		3	1				18							3
AAF00925	Intestinal fatty acid binding protein (*D. rerio*)	(X) 1.7E-58	131 (83.9%)	11_1							6							
BAB40965	28 kDa-1e apolipoprotein[B] (*A. japonica*)[d]	(X) 9.7E-55	254 (46.0%)	11_1							6							
n/a[c]	unknown	n/a	n/a	11_3							2				2			1
AAG60018	Acidic mammalian chitinase (*Mus musculus*)	(X) 7.5E-7	51 (45.0%)	11_3							1							
AAM73701	C1q-like adipose specific protein (*S. fontinalis*)	(X) 1.1E-31	162 (45%)	11_4								3		1	1			1
BG935597	unknown liver clone SL1-0959 (*S. salar*)	(N) 2.0E-12	115 (84.3%)	10					1							2		
CAA65953	β-globin (*S. salar*)	(X) 1.4E-82	147 (99.3%)	10			1				1							
CAA49678	Trypsin II (*S. salar*)	(X) 3.7E-141	231 (100%)	10_1							3				4			2
n/a[c]	unknown	n/a	n/a	10_1							1							1
AJ424208	unknown kidney clone k04A08 (*S. salar*)	(N) 2.1E-53	336 (83.6%)	10_2		1												
AAD30275	Heat shock protein hsp90 β (*S. salar*)	(X) 0	369 (100%)	10_2	2	2	5		3		4	1	1	1	3	13	3_1	
AAF61069	Galectin (*P. olivaceus*)	(X) 8.1E-27	129 (44.1%)	10_2														1
P07514	NADH-cytochrome b5 reductase (*Bos taurus*)	(X) 7.2E-105	267 (66.2%)	10_3									1		5			1
AAF89686	Catalase (*D. rerio*)	(X) 4.2E-38	116 (73.2%)	10_3											3			
AAD34044	CGI-49 protein (*H. sapiens*)	(X) 8.4E-48	139 (61.4%)	10_3							3							
NP_571833	Stomatin (*D. rerio*)	(X) 1.2E-100	281 (70.8%)	10_3											1			
BAB40965	28 kDa-1e apolipoprotein[C] (*A. japonica*)[d]	(X) 2.6E-55	254 (44.4%)	10_4						1	4							1
AAL40376	High choriolytic enzyme 1 (*Takifugu rubripes*)	(X) 4.1E-84	215 (67.4%)	10_4							4							4
CAC21508	Meprin A α (*H. sapiens*)	(X) 1.4E-32	122 (53.2%)	9														
CAC87888	Toad pancreatic chitinase (*Bufo japonicus*)	(X) 0	327 (51.6%)	9_2							16							
AAC24501	GDP-D-mannose-4,6-dehydratase (*H. sapiens*)	(X) 6.4E-51	121 (76.8%)	9_2							1				1			
P02593	Calmodulin (*H. sapiens*)	(X) 7.9E-82	149 (100%)	9_2	2		5		2		4	3			2			

[a]Compiled using the August 16, 2003 version of the GRASP EST database, and excluding mitochondrial and ribosomal EST clusters. ESTs assembled using PHRAP.
[b]For notes on *S. salar* libraries and library groups, see Table 2. Library abbreviations follow: pc indicates all pyloric caecum libraries; br, brain; es, esophagus; gi, gill; he, normalized heart; ki, all kidney libraries; li, normalized liver; mg, all mixed gut (stomach + mid-gut + hind-gut) libraries; ov, ovary; pi, pituitary gland; re, retina; skb, normalized (spleen + kidney + brain); sp, all spleen libraries; te, testis and w, all whole juvenile libraries. Subscripts following EST numbers denote the number of reverse or duplicate forward reads matching forward reads in the cluster.
[c]See notes on Table 3.
[d]Paralogs discussed in text. The aligned translations of B and C differ at 13 of 228 residues (94.3% identity); each differs from the translation of A at 104 of 228 residues (54.4% identity).

Table 6. Analysis of Cross-Species Hybridization to 7356 Element Salmonid cDNA Microarray[a]

	Target hybridized to array[a]	Atlantic salmon (AS)[b]		Rainbow trout (RT)[b]		Lake whitefish (LW)[b]		Rainbow smelt (S)[b]	
	Elements printed on array[a]	AS (6440)	RT (916)	AS (6440)	RT (916)	AS (6440)	RT (916)	AS (6440)	RT (916)
Slide 1 AS: CL010.015 RT: CL010.018 LW: CL010.021 S: CL010.024	threshold:[c]	381.1	381.1	362.2	362.2	290.0	290.0	331.4	331.4
	elements passing threshold (%):	4581 (71.1%)	512 (55.9%)	4488 (69.7%)	535 (58.4%)	4633 (71.9%)	510 (55.7%)	2027 (31.5%)	240 (26.2%)
	signal total:[d]	2.00E7	1.96E6	1.88E7	2.44E6	1.50E7	1.87E6	5.68E6	6.80E5
	signal mean:[d]	3111.1	2141.9	2914.7	2666.9	2328.9	2038.4	882.5	742.7
	background total:[d]	1.68E6	2.34E5	1.56E6	2.19E5	1.24E6	1.73E5	1.38E6	1.93E5
	background mean:[d]	261.2	255.4	242.4	238.6	193.0	188.4	214.2	210.3
	signal/background[e]	11.9	8.4	12.0	11.2	12.1	10.8	4.1	3.5
Slide 2 AS: CL010.016 RT: CL010.019 LW: CL010.022 S: CL010.025	threshold:[c]	386.7	386.7	362.3	362.3	294.3	294.3	288.0	288.0
	elements passing threshold (%):	4479 (69.5%)	486 (53.1%)	4496 (69.8%)	539 (58.8%)	4674 (72.6%)	520 (56.8%)	2881 (44.7%)	302 (33.0%)
	signal total:[d]	1.92E7	1.82E6	1.90E7	2.56E6	1.52E7	1.86E6	7.54E6	9.36E5
	signal mean:[d]	2981.5	1991.7	2950.5	2796.6	2363.9	2029.1	1171.4	1022.3
	background total:[d]	1.70E6	2.36E5	1.55E6	2.17E5	1.27E6	1.76E5	1.19E6	1.67E5
	background mean:[d]	263.5	257.8	239.9	236.7	196.8	192.2	184.6	182.1
	signal/background[e]	11.3	7.7	12.3	11.8	12.0	10.6	6.3	5.6
Slide 3 AS: CL010.017 RT: CL010.020 LW: CL010.023 S: CL010.026	threshold:[c]	366.8	366.8	308.6	308.6	312.4	312.4	317.9	317.9
	elements passing threshold (%):	4544 (70.6%)	480 (52.4%)	5151 (80.0%)	618 (67.5%)	4657 (72.3%)	528 (57.6%)	2451 (38.1%)	275 (30.0%)
	signal total:[d]	2.09E7	2.00E6	1.87E7	2.57E6	1.61E7	1.91E6	6.59E6	8.55E5
	signal mean:[d]	3249.8	2179.3	2910.6	2810.7	2497.7	2086.9	1023.9	933.9
	background total:[d]	1.60E6	2.21E5	1.35E6	1.88E5	1.33E6	1.85E5	1.34E6	1.88E5
	background mean:[d]	248.4	241.0	209.5	205.3	206.9	201.7	208.4	204.8
	signal/background[e]	13.1	9.0	13.9	13.7	12.1	10.4	4.9	4.6
OVERALL	mean no. (%) passing threshold:	4535 (70.4%)	493 (53.8%)	4712 (73.2%)	564 (61.6%)	4655 (72.3%)	519 (56.7%)	2453 (38.1%)	272 (29.7%)
	mean total signal:[d]	2.01E7	1.93E6	1.88E7	2.53E6	1.54E7	1.88E6	6.61E6	8.24E5
	SEM:	4.99E5	5.25E4	8.16E4	4.19E4	3.31E5	1.64E4	5.37E5	7.56E4
	mean total background:[d]	1.66E6	2.30E5	1.48E6	2.08E5	1.28E6	1.78E5	1.30E6	1.82E5
	SEM:	3.03E4	4.80E3	6.81E4	9.89E3	2.68E4	3.63E3	5.82E4	7.90E3

[a]The 7356 elements (spots) on this microarray include 3119 unique *S. salar* cDNAs (3018 spotted in duplicate, 101 in quadruplicate) and 438 unique *O. mykiss* cDNAs (418 spotted in duplicate, 20 in quadruplicate). All targets were Cy3-labeled and hybridized to microarray slides from the same batch (CL010).
[b]Atlantic salmon = *Salmo salar;* rainbow trout = *Oncorhynchus mykiss;* lake whitefish = *Coregonus clupeaformis;* rainbow smelt = *Osmerus mordax.*
[c]Threshold, used to determine "elements present," was calculated using raw median signal values from 1356 control elements: 204 buffer (3 × SSC) alone, 912 bare glass, and 240 green fluorescent protein (GFP) cDNA included on the chip (but not used in this study) for gridding purposes. Threshold = mean + 2 SD.
[d]All slides in this study were scanned at 90% laser power and PMT set to 75. Median intensity values (raw signal or background) were used for analyses.
[e]Signal/background calculated as raw (not background-corrected) total signal divided by raw total background.



Labeled target	n	x̄ total signal on AS spots	AS signal rel. to smelt	x̄ total signal on RT spots	RT signal rel. to smelt	genome size (pg/N)	genome size rel. to smelt
AS	3	2.01E7	3.04	1.93E6	2.34	3.10	5.00
RT	3	1.88E7	2.85	2.53E6	3.07	2.59	4.18
LW	3	1.54E7	2.34	1.88E6	2.28	3.44	5.55
smelt	3	0.661E7	1.00	0.824E6	1.00	0.62	1.00

Figure 2 Evolutionary relationships, genome sizes, and microarray hybridization characteristics of three salmonids relative to smelt. (*A*) Phylogenetic tree, based on morphological characters, showing evolutionary relationships among teleosts relevant to this study, and other fish orders with genome projects (Nelson 1994). (*B*) Phylogenetic tree, based on morphological characters, showing evolutionary relationships of select salmonids (Smith and Stearley 1989; Kido et al. 1991). *Arrows* indicate putative genome duplication events (Wolfe 2001). (*C*, *D*) Mean total signals on Atlantic salmon (AS) or rainbow trout (RT) chip elements/spots (Table 6) are converted to "smelt units" by dividing by 0.661E7 for AS chip elements, or 0.824E6 for RT chip elements. Genome sizes for AS (*Salmo salar*), RT (*Oncorhynchus mykiss*), and smelt (*Osmerus eperlanus*, close relative of *Osmerus mordax* used in this study) were measured by DNA flow cytometry (Vinogradov 1998). Genome size of lake whitefish (LW, *Coregonus clupeaformis*) was measured by Feulgen densitometry (Booke 1968). Error bars (*C*) show mean total signal SEM values (Table 6) converted to "smelt units" as above. n indicates number of microarrays hybridized with labeled target from each species.

lowed by AS (mean of three slides: 1.93E6, SEM 5.25E4), LW (mean of three slides: 1.88E6, SEM 1.64E4), and rainbow smelt (mean of three slides: 8.24E5, SEM 7.56E4; Table 6).

The ranking of hybridization performances conformed to expectations, given the evolutionary relationships of the species tested (Fig. 2A,B). AS and RT, members of the subfamily Salmoninae, diverged in the Miocene 8 to 20 million years ago (Stearley 1992; Devlin 1993; Coe et al. 1995). Phylogenies based on morphological (Nelson 1994) and molecular (Phillips and Oakley 1997) data show that the genera *Salmo* and *Oncorhynchus* are more closely related to one another than either group is related to *Coregonus*, the genus of LW. On both AS and RT chip elements, hybridization performance of LW target ranks third behind AS and RT targets (Table 6, Fig. 2C,D). Because the mean numbers of AS and RT elements passing threshold are comparable for AS, RT, and LW targets (Table 6), the lower signal from LW-hybridized slides likely reflects lower percentage of identity between salmonine probe and coregonine target sequences. These hybridization results match predicted distances of divergence for the salmonid species tested (Fig. 2B). Our preliminary analysis of AS and RT putatively orthologous EST contigs (primarily 3′) shows ~94% identity and is in agreement with the success of these species' targets on one anothers' probes (Table 6, Fig. 2C,D). Our EST database does not yet contain adequate numbers of LW EST contigs to permit large-scale alignment of putative orthologous sequences. However, the high performances of LW targets on AS and RT probes (Table 6, Fig. 2C,D) are suggestive of high similarity between LW and salmonine orthologous cDNAs. Hybridization performances of rainbow smelt targets were less than half those of salmonid (AS, RT, or LW) targets (Table 6, Fig. 2C,D), likely due to lower similarity (reflecting longer time since divergence) between cDNAs from members of the order Salmoniformes and orthologous sequences from members of the order Osmeriformes (Fig. 2A).

Identification of Candidate Duplicated Genes

Osmerids are diploid and salmonids are degenerate tetraploids (Ohno et al. 1968; Fig. 2C,D), placing the putative, salmonid-specific genome duplication event after the divergence of Osmeridae and Salmonidae (Fig. 2A). Because at least 50% of recent gene duplicates are thought to persist in salmonids (Bailey et al. 1978), it is expected that gene family expansion (the presence of multiple expressed paralogs) would be widespread in this group. Preliminary comparisons of robust EST clusters (in single AS libraries/library groups) that have identical top BLASTX hits reveal the presence of multiple distinct forms (not splice variants) of several genes (i.e., novel member of chitinase family, β-globin, and serum lectin, Table 3; 28 kD − 1e apolipoprotein, Table 5). Further work (i.e., molecular phylogenetics, fluorescence in situ hybridization) will be required to distinguish paralogs arising during the recent salmon-specific genome duplication from those with origins in other gene/genome duplication events. The GRASP EST database, and an improved salmonid presence in GenBank databases, will facilitate identification of additional members in gene families, contributing to a better understanding of the evolution of related genes within and between genomes.

METHODS

Aquaculture and Sampling

S. salar (McConnell strain) juveniles were obtained from Heritage Aquaculture (British Columbia, Canada), and cultured throughout their life history. Subadult *S. salar* were sampled from various tissues at 2.75 years of age (Fisheries and Oceans Canada, West Vancouver, British Columbia) and used for generating all adult cDNA libraries and labeled targets for microarray hybridizations. For juvenile cDNA libraries, *S. salar* (McConnell strain) and *O. mykiss* (Tzenzaicut Lake strain) were obtained from SeaSpring Hatchery (Duncan, British Columbia) and Vancouver Island Trout Hatchery (Duncan, British Columbia), respectively. For labeled targets used in microarray hybridizations, embryonic stages of *O. mykiss* were derived from a domesticated strain (Spring Valley Trout Farm, Langley, British Columbia) and cultured to ~80 g before sampling. *O. mykiss* gonadal tissues (≥1.5 years; Spring Valley Strain), used to generate subtractive cDNA libraries, were obtained from Mountain Trout Sales (Sooke, British Columbia). *O. tshawytscha* tissues were obtained from 4-year-old females (Robertson Creek, British Columbia); *O. nerka* tissues were obtained from whole juvenile fish (Dr. L.J. Albright, Simon Fraser University); *C. clupeaformis* brain and liver were obtained from 3-year-old animals (Laboratoire Bernatchez, Université Laval, Quebec), and *Osmerus mordax* livers were obtained from adult smelt (NRC Institute for Marine Biosciences).

Fish were raised in fiberglass tanks with natural lighting and at densities <10 kg/m^3 with water input rate >1 L min^{-1} kg^{-1}. *S. salar* and *O. tshawytscha* were reared in fresh 10°C well water until smolt stage (1.5 years) and then transferred to sea water until sexual maturation. *O. mykiss* were cultured only in fresh 10°C well water. Most fish were fed to satiation three times per day with commercial salmon diets (Pacific Apollo 1000, Moore Clarke, Vancouver, British Columbia) comprised of 40% protein and 25% lipid.

Fish were killed by a blow to the head, followed by rapid dissection. Tissues were flash-frozen in liquid nitrogen and stored at −80°C until RNA extraction. For gut tissues, discrete sections were excised and the lumen gently rinsed free of food and feces with a stream of ice-cold phosphate-buffered saline (10 mM PO_4, 138 mM NaCl, and 27 mM KCl at pH 7.4).

cDNA Libraries

Flash-frozen tissues were ground by using baked (5 h, 220°C) mortars and pestles under liquid N_2, and poly(A)+ RNA was purified by using MicroPoly(A)Pure kits (Ambion) or Oligotex Direct mRNA Micro Kits (Qiagen). With the exception of the *O. nerka* libraries, the normalized *S. salar* mixed tissue library, and the suppression subtractive hybridization (SSH) libraries, cDNA libraries were directionally constructed (5′ EcoRI, 3′ XhoI), using the pBluescript II XR cDNA Library Construction Kit, following manufacturer's instructions (Stratagene). Size fractionation, performed on XhoI-digested cDNAs immediately prior to ligation into vector, was by 1% agarose gel extraction (Qiagen). *O. nerka* libraries were size-fractionated by using CHROMA Spin-400 columns (Clontech), and directionally constructed (5′ SfiIA, 3′ SfiIB) in pDNR-LIB using the Creator SMART cDNA Library Construction Kit (Clontech). Select cDNA libraries were normalized to Cot = 5 by using the Soares method (Soares et al. 1994; Bonaldo et al. 1996). The normalized (Cot = 10) *S. salar* mixed tissue (spleen, head kidney, brain) library was directionally constructed in pCMV Sport6.1 (ResGen). SSH libraries were constructed by using the PCR-Select cDNA Subtraction Kit (Clontech) following manufacturer's instructions, and were TA cloned into pCR 4-TOPO (Invitrogen). Insert sizes of cDNA libraries were determined by visual comparison of clone restriction fragments with the DNA size markers λHindIII (GIBCO-BRL) and 1-kb ladder (GIBCO-BRL).

Sequencing, Sequence Analysis, and Contig Assembly

Libraries were manually arrayed in 96-well microtiter plates or were robotically arrayed in 384-well plates. Glycerol stocks of overnight cultures were prepared in 96-well or 384-well format. Plasmid DNAs were extracted and BigDye Terminator (ABI) cycle sequenced on ABI 3700 and 377 sequencers by using conventional procedures and the following primers: 5′-T_{18}-3′, M13 forward (5′-GTAAAACGACGGCCAGT-3′), and M13 reverse (5′-AACAGCTATGACCATG-3′ or 5′-AACAGCTATGACCAT-3′). Basecalling from chromatogram traces was performed by using PHRED (Ewing and Green 1998; Ewing et al. 1998). Vector, poly-A tails, and low-quality regions were trimmed from EST sequences; sequences that had <100 good-quality bases after trimming were discarded.

Vectors were screened by using cross_match (part of the PHRAP package, version 0.990329), with minscore = 18. This is more sensitive than Consed, allowing detection of adaptor sequences in subtractive libraries. All vector was trimmed from the ends of the sequence. If there was remaining vector in the middle, it was removed and the shorter of the two remaining fragments trimmed with it.

To trim poly-A tails, sequences were scanned from their ends forward to the beginning of the last run of consecutive As. If the tail of the sequence up to that point was at least 60% A, then it was considered part of the tail. This test was repeated from that point forward until it failed. The portion of the sequence that passed was considered poly-A tail. If this test found nothing, then the last 100 bases of the sequence were scanned for a run of at least 15 consecutive As. If found, then the trailing sequence was assumed to be bad or vector, and all sequence up to and including the run of As was trimmed. To scan for poly-T tails, the same tests were performed on reverse-complemented sequences. Sequences were not considered poly-A or poly-T tails if they were <10 bases in length.

PHRAP (http://www.genome.washington.edu/UWGC), under stringent clustering parameters (minimum score, 100; repeat stringency, 0.99), was used to assemble ESTs into contigs. Contig consensus sequences and singleton sequences were aligned with nonredundant GenBank nucleotide and amino acid sequence databases by using BLASTN and BLASTX, respectively (Altschul et al. 1990, 1997). Results of EST clustering using CAP3 (40-bp overlap, 95% identity, other parameters default) and stackPACK (using RepeatMasked sequences without quality scores) are available at the GRASP Web site (http://web.uvic.ca/cbr/grasp). To determine the approximate amount of ribosomal and mitochondrial sequence in the GRASP EST database, each species' ESTs were aligned against a BLAST database containing the same species' GenBank sequences annotated as ribosomal plus the GenBank mitochondrial sequences from that species or its closest relative. BLAST hits with E values <10^{-5} qualified ESTs as ribosomal or mitochondrial.

Assembled EST contigs were scanned for repeats by using REPuter (Kurtz et al. 2001). Candidate repeats (length >50 bases, fewer than eight mismatches, and E < 10^{-4}) were assembled into contigs by using PHRAP, and compared with GenBank nr and nt databases by using BLASTX and BLASTN, respectively. Threshold for a significant BLAST hit was set at 10^{-15}. BLAST results were deposited in a database, and a Web interface for querying was implemented (http://woodstock.ceh.uvic.ca/nkuipers/public_html/).

O. mykiss orthologs to *S. salar* contigs were detected by semiglobal (end-gaps-free) pairwise alignment of forward and reverse-complement contigs. Alignments with overlaps of <100 nucleotides were discarded. *O. mykiss* contigs were considered orthologous to an *S. salar* contig if either the forward- or reverse-complement alignment showed at least 80% identity.

Functional Characterization of EST Contigs

By using the March 3, 2003, version of the GRASP EST database, assembled *S. salar* ESTs from select organ-specific libraries or library groups (pyloric caecum, gill, mixed gut, ovary, and pituitary gland), and all *S. salar* libraries collectively, were compared via BLASTX with annotated protein sequences from the GO database (November 2002 version; Table 4). Sequences with significant matches (E-value < 10^{-5}) were classified according to the GO classification(s) of their strongest hit. For several GO functional categories of genes, Z-statistics were used to determine if there were significant differences between the proportions of assembled ESTs in an organ-specific library/library group (i.e., gill) and the proportions of assembled ESTs in remaining (i.e., non-gill) libraries. Z-statistics, used for the comparison of two sample

proportions (Anderson and Finn 1997), were calculated by using the following equation:

$$Z = \frac{p1 - p2}{\text{Square root } (p^*(1 - p)^*(1/nl + 1/n2))}$$

where p1 is the proportion of assembled ESTs in the organ of interest (2/444 for gill library, antioxidant GO category; Table 4), p2 is the proportion of assembled ESTs in nonorgan libraries (9/6869 for nongill, antioxidant; see Supplemental data at http://web.uvic.ca/cbr/grasp), p is the overall proportion (10/6937 for antioxidant; Table 4), n1 is the number of organ-specific assembled ESTs (444 for gill; Table 4), and n2 is the number of nonorgan assembled ESTs (6869 for nongill; see Supplemental data at http://web.uvic.ca/cbr/grasp). Z has a standard normal distribution, so *P*-values are computed as $1 - (\text{CDF}(\text{abs}(Z)) \times 2)$, where CDF is the cumulative distribution function of the standard normal distribution and abs is absolute value. This *P*-value gives a two-tailed test for the probability that the proportions of organ and nonorgan EST contigs in a given molecular function category are equal.

Microarray Fabrication and Quality Control

The 3557 clones from 18 high-complexity salmonid cDNA libraries/library groups (Table 2) were selected with an emphasis on immune relevant genes. Clones were robotically rearrayed from daughter glycerol stock 384-well plates into 96-well plates prefilled with 7% glycerol in LB + ampicillin, incubated overnight at 37°C, and checked for uniform optical density. Plasmid inserts were PCR-amplified in a Tetrad PTC-200 thermocycler (MJ Research) by using 1 µL overnight culture, 0.2 µM M13/pUC forward primer (5′-CCCAGTCACGACGTTGTAAAACG-3′), 0.2 µM M13/pUC reverse primer (5′-AGCGGATAACAATTTCACACAGG-3′), 2 mM $MgCl_2$, 10 mM Tris-HCl, 50 mM KCl, 250 µM dNTPs, 1U AmpliTaq (PerkinElmer), and nuclease-free H_2O (GIBCO) to 100 µL. PCR conditions were as follows: 2 min at 95°C denaturation; 35 cycles of 30 sec at 95°C, 45 sec at 60°C, and 3 min at 72°C; and 7 min at 72°C. Five microliters of each PCR product were run on a 1% agarose gel to assess yield and quality. Out of 3557 clones, there were 3312 strong single bands (93%), 170 absent (5%), and 75 multiple bands (2%). PCR products were robotically cleaned (Qiagen) and consolidated into 384-well plates, lyophilized by speed-Vac, and resuspended in 15 µL 3× SSC.

All cDNAs were printed as double, side-by-side spots on Telechem Superamine slides (Arrayit) with the Biorobotics Microgrid II microarray printer (Apogent Discoveries). Microspot 10K quill pins (Biorobotics) in a 48-pin tool were used to deposit ~0.5 nL (0.2 ng cDNA) per spot onto the slide. The resulting microarrays have a 4 × 12 subgrid layout with 132 spots per subgrid, each spot having approximate diameter and pitch of 100 µm and 250 µm, respectively. A 280-bp GFP (green fluorescent protein) cDNA was amplified from a GFP clone (Clontech) by using the primers (5′-GAAACATTCTTGGACACAAATTGG-3′) and (5′-GCAGCTGTTACAAACTCAAGAAGG-3′), and printed in subgrid corners to assist in placing on the grid. The slides were crosslinked in a UV Stratalinker 2400 (Stratagene) at 120 mJ. Spot morphology was assessed by visual inspection, SYBR Green 1 (Molecular Probes) staining, or hybridization with labeled nonspecific probe. To check clone tracking, 42 high-quality sequences were obtained from randomly selected wells of the cleaned, consolidated 384-well plates used for microarray printing. All 42 had BLAST identifiers matching gene identifications predicted from the rearray spreadsheet, indicating highly accurate clone tracking throughout the process of microarray fabrication.

Microarray Hybridization and Analysis

This microarray experiment was designed to comply with MIAME guidelines (Brazma et al. 2001). All scanned microarray TIF images, an ImaGene grid, the gene identification file, and ImaGene quantified data files are available at http://web.uvic.ca/cbr/grasp. To minimize technical variability, all targets were synthesized in one round, and all hybridizations were conducted simultaneously on slides from a single batch (CL010, Table 6). Total RNA, prepared from flash-frozen adult liver tissues using TRIzol reagent and methods (Invitrogen), was quantified and quality-checked by spectrophotometer and agarose gel. Hybridizations were performed by using the Genisphere Array50 kit and instructions. Briefly, 15 µg total RNA were reverse-transcribed by using a special oligo d(T) primer with a 5′ unique sequence overhang for the Cy3 labeling reactions. Microarrays were prepared for hybridization by washing two times at 5 min in 0.1% SDS, washing five times at 1 min in MilliQ H_2O, immersing 3 min in 95°C MilliQ H_2O, and drying by centrifugation (5 min at 2000 rpm in 50-mL conical tube). The cDNA was hybridized to the salmon cDNA microarray in a formamide-based buffer (25% formamide, 4× SSC, 0.5% SDS, 2× Denhardt's solution) for 16 h at 48°C. The arrays were washed one time for 10 min at 48°C (2× SSC, 0.1% SDS), two times for 5 min in 2× SSC, 0.1% SDS at room temperature (RT), two times for 5 min in 1× SSC at RT, and two times for 5 min in 0.1× SSC at RT, and dried by centrifugation. The Cy3 3-dimensional fluorescent molecules (3DNA capture reagent, Genisphere) were hybridized to the bound cDNA on the microarray; the Cy3 3DNA capture reagent bound to its complementary cDNA capture sequence on the Cy3 oligo d(T) primer. The second hybridization was done for 3 h at 48°C, and washed and dried as before.

The fluorescent images of hybridized arrays were acquired by using ScanArray Express (PerkinElmer). The Cy3 cyanine fluor was excited at 543 nm, and the same laser power (90%) and photomultiplier tube (PMT) setting (75) were used for all slides in the study. Fluorescent intensity data was extracted by using Imagene 5.5 software (Biodiscovery). To avoid transformations associated with background correction (i.e., setting negative background corrected median signal values to zero), raw median signal values were analyzed. No normalization was applied to the data. From the raw Imagene fluorescence intensity report files, the gene lists were sorted, and median signal values from 1356 control elements (204 buffer alone, 912 bare glass, and 240 GFP cDNA) were analyzed. For each slide, threshold was calculated as the mean intensity for these 1356 controls plus 2 SD. For data analyses, the 6440 *S. salar* (AS) chip elements and 916 *O. mykiss* (RT) chip elements were considered separately. The mean numbers of AS and RT elements passing threshold, mean total slide signal (salmonid elements only) and SEM, mean total slide background (local background fluorescence intensities associated with salmonid elements) and SEM, and average signal and background per salmonid element were calculated by slide and by species. To assess array-wide performance, signal-to-background ratio was calculated as raw total signal divided by raw total background.

ACKNOWLEDGMENTS

This research was supported by Genome Canada, Genome BC, and the Province of BC and, additionally, by the Natural Sciences and Engineering Research Council of Canada (B.K., W.D.). We would like to thank Carlo Biagi, Steve Dann, and Shelby Temple for their assistance in obtaining tissues for cDNA library construction; Bento Soares and Brian Berger for providing methods and advice on normalizing cDNA libraries, and all those at the BCCA Genome Sciences Centre who contributed to this work.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Allendorf, F.W. and Danzmann, R.G. 1997. Secondary tetrasomic segregation of MDH-B and preferential pairing of homologues in rainbow trout. *Genetics* **145:** 1083–1092.

Allendorf, F.W. and Thorgaard, G.H. 1984. Tetraploidy and the evolution of salmonid fishes. In *Evolutionary genetics of fishes* (ed. B.J. Turner), pp. 1–53. Plenum Press, New York.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller,

W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.
Anderson, T.W. and Finn, J.D. 1996. *The new statistical analysis of data*. Springer-Verlag, New York.
Bailey, G.S., Poulter, R.T., and Stockwell, P.A. 1978. Gene duplication in tetraploid fish: Model for gene silencing at unlinked duplicated loci. *Proc. Natl. Acad. Sci.* **75:** 5575–5579.
Bailey, G.S., Williams, D.E., and Hendricks, J.D. 1996. Fish models for environmental carcinogenesis: The rainbow trout. *Environ. Health Perspect.* **104:** 5–21.
Basu, N., Todgham, A.E., Ackerman, P.A., Bibeau, M.R., Nakano, K., Schulte, P.M., and Iwama, G.K. 2002. Heat shock protein genes and their functional significance in fish. *Gene* **295:** 173–183.
Bonaldo, M.F., Lennon, G., and Soares, M.B. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6:** 791–806.
Booke, H.E. 1968. Cytotaxonomic studies of the coregonine fishes of the Great Lakes, USA: DNA and karyotype analysis. *J. Fish. Res. Board Can.* **25:** 1667–1687.
Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al. 2001. Minimum information about a microarray experiment (MIAME): Toward standards for microarray data. *Nat. Genet.* **29:** 365–371.
Buddington, R.K. and Diamond, J.M. 1986. Aristotle revisited: The function of pyloric caeca in fish. *Proc. Natl. Acad. Sci.* **83:** 8012–8014.
Burk, R.F., Hill, K.E., and Motley, A.K. 2003. Selenoprotein metabolism and function: Evidence for more than one function for selenoprotein P. *J. Nutr.* **133:** 1517S–1520S.
Coe, I.R., von Schalburg, K.R., and Sherwood, N.M. 1995. Characterization of the Pacific salmon gonadotropin-releasing hormone gene, copy number, and transcription start site. *Mol. Cell. Endocrinol.* **115:** 113–122.
Danielson, P.B. 2002. The cytochrome p450 superfamily: Biochemistry, evolution and drug metabolism in humans. *Curr. Drug Metab.* **3:** 561–597.
Davey, G.C., Caplice, N.C., Martin, S.A., and Powell, R. 2001. A survey of genes in the Atlantic salmon (*Salmo salar*) as identified by expressed sequence tags. *Gene* **263:** 121–130.
Deplancke, B. and Gaskins, H.R. 2002. Redox control of the transsulfuration and glutathione biosynthesis pathways. *Curr. Opin. Clin. Nutr. Metab. Care* **5:** 85–92.
Devlin, R.H. 1993. Sequence of sockeye salmon type 1 and type 2 growth hormone genes and the relationship of rainbow trout with Atlantic and Pacific salmon. *Can. J. Fish. Aquat. Sci.* **50:** 1738–1748.
Devlin, R.H., Biagi, C.A., Yesaki, T.Y., Smailus, D.E., and Byatt, J.C. 2001. Growth of domesticated transgenic fish. *Nature* **409:** 781–782.
Douglas, S.E. and Gallant, J.W. 1998. Isolation of cDNAs for trypsinogen from the winter flounder. *Pleuronectes americanus*. *J. Mar. Biotechnol.* **6:** 214–219.
Douglas, S.E., Bullerwell, C.E., and Gallant, J.W. 1999a. Molecular investigation of aminopeptidase N expression in the winter flounder. *Pleuronectes americanus*. *J. Appl. Ichthyol.* **15:** 80–86.
Douglas, S.E., Gallant, J.W., Bullerwell, C.E., Wolff, C., Munholland, J., and Reith, M.E. 1999b. Winter flounder expressed sequence tags: Establishment of an EST database and identification of novel fish genes. *Mar. Biotechnol.* **1:** 458–464.
Eshel, R., Besser, M., Zanin, A., Sagi-Assif, O., and Witz, I.P. 2001. The FX enzyme is a functional component of lymphocyte activation. *Cell. Immunol.* **213:** 141–148.
Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using PHRED, II: Error probabilities. *Genome Res.* **8:** 186–194.
Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using PHRED, I: Accuracy assessment. *Genome Res.* **8:** 175–185.
Faillace, M.P., Julian, D., and Korenbrot, J.I. 2002. Mitotic activation of proliferative cells in the inner nuclear layer of the mature fish retina: Regulatory signals and molecular markers. *J. Comp. Neurol.* **451:** 127–141.
Gregory, T.R. 2002. Animal genome size database. http://www.genomesize.com.
Holland, P.W., Garcia-Fernandez, J., Williams, N.A., and Sidow, A. 1994. Gene duplications and the origins of vertebrate development. *Development (Suppl.) 1994:* 125–133.
Katchamart, S., Miranda, C.L., Henderson, M.C., Pereira, C.B., and Buhler, D.R. 2002. Effect of xenoestrogen exposure on the expression of cytochrome P450 isoforms in rainbow trout liver. *Environ. Toxicol. Chem.* **11:** 2445–2451.
Kido, Y., Aono, M., Yamaki, T., Matsumoto, K., Murata, S., Saneyoshi, M., and Okada, N. 1991. Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution. *Proc. Natl. Acad. Sci.* **88:** 2326–2330.
Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. 2001. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **29:** 4633–4642.
Madigou, T., Uzbekova, S., Lareyre, J.J., and Kah, O. 2002. Two messenger RNA isoforms of the gonadotrophin-releasing hormone receptor, generated by alternative splicing and/or promoter usage, are differentially expressed in rainbow trout gonads during gametogenesis. *Mol. Reprod. Dev.* **63:** 151–160.
Nelson, J.S. 1994. *Fishes of the world*, 3rd ed. John Wiley & Sons, New York.
Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlang, Heidelberg, Germany.
Ohno, S., Stenius, C., Faisst, E., and Zenzes, M.T. 1965. Post-zygotic chromosomal rearrangements in rainbow trout (*Salmo irideus* GIBBONS) *Cytogenetics* **4:** 117–129.
Ohno, S., Wolf, U., and Atkin, N.B. 1968. Evolution from fish to mammals by gene duplication. *Hereditas* **59:** 169–187.
Ohyama, C., Smith, P.L., Angata, K., Fukuda, M.N., Lowe, J.B., and Fukuda, M. 1998. Molecular cloning and expression of GDP-D-mannose-4,6-dehydratase, a key enzyme for fucose metabolism defective in Lec13 cells. *J. Biol. Chem.* **273:** 14582–14587.
Phillips, R.B. and Oakley, T.H. 1997. Phylogenetic relationships among the Salmoninae based on nuclear and mitochondrial DNA sequences. In *Molecular systematics of fishes* (eds. T.D. Kocher and C.A. Stepien), pp. 145–162. Academic Press, San Diego, CA.
Phillips, R. and Ráb, P. 2001. Chromosome evolution in the Salmonidae (Pisces): An update. *Biol. Rev.* **76:** 1–25.
Rabinovich, G.A., Rubinstein, N., and Toscano, M.A. 2002. Role of galectins in inflammatory and immunomodulatory processes. *Biochim. Biophys. Acta* **1572:** 274–284.
Schomburg, L., Schweizer, U., Holtmann, B., Flohe, L., Sendtner, M., and Kohrle, J. 2003. Gene disruption discloses role of selenoprotein P in selenium delivery to target tissues. *Biochem. J.* **370:** 397–402.
Shum, B.P., Guethlein, L., Flodin, L.R., Adkison, M.A., Hedrick, R.P., Nehring, R.B., Stet, R.J.M., Secombes, C., and Parham, P. 2001. Modes of salmonid MHC class I and II evolution differ from the primate paradigm. *J. Immunol.* **166:** 3297–3308.
Sidow, A. 1996. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6:** 715–722.
Smith, G.R. and Stearley, R.F. 1989. The classification and scientific names of rainbow and cutthroat trouts. *Fisheries* **14:** 4–10.
Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.
Stearley, R.F. 1992. Historical ecology of the Salmoninae, with special reference to *Oncorhynchus*. In *Systematic historical ecology and North American freshwater fishes* (ed. R.L. Mayden), pp. 622–658. Stanford University Press, Stanford, CA.
Tipsmarck, C.K., Madsen, S.S., Seidelin, M., Christensen, A.S., Cutler, C.P., and Cramb, G. 2002. Dynamics of Na^+,K^+,$2Cl^-$ cotransporter and Na^+,K^+-ATPase expression in the branchial epithelium of brown trout (*Salmo trutta*) and Atlantic salmon (*Salmo salar*). *J. Exp. Zool.* **293:** 106–118.
Vinogradov, A.E. 1998. Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. *Cytometry* **31:** 100–109.
Wolfe, K.H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2:** 333–341.
Zhang, C., Brown, S.B., and Hara, T.J. 2001. Biochemical and physiological evidence that bile acids produced and released by lake char (*Salvelinus namaycush*) function as chemical signals. *J. Comp. Physiol. B* **171:** 161–171.

WEB SITE REFERENCES

http://www.geneontology.org; the Gene Ontology Consortium (2001).
http://web.uvic.ca/cbr/grasp; University of Victoria Centre for Biomedical Research.
http://woodstock.ceh.uvic.ca/nkuipers/public_html/; Web interface for querying a database containing BLAST-identified candidate repeats in the GRASP EST database.
http://www.genome.washington.edu/UWGC; University of Washington Genome Centre (PHRED version 0.990722.j; PHRAP version 0.990329).

Received June 25, 2003; accepted in revised form December 12, 2003.