Sequencing the genome of the Atlantic salmon (Salmo salar)

William S Davidson*, Ben F Koop1, Steven JM Jones1, Patricia Iturra4, Rodrigo Vidal1, Alejandro Maass6, Inge Jonassen7, Sigbjorn Lien8 and Stig W Omholt8

Abstract
The International Collaboration to Sequence the Atlantic Salmon Genome (ICSASG) will produce a genome sequence that identifies and physically maps all genes in the Atlantic salmon genome and acts as a reference sequence for other salmonids.

Economic, societal and scientific importance of salmonids
The family Salmonidae comprises 11 genera and includes salmon, trout, char, freshwater whitefishes, ciscos and graylings [1]. Many salmonid species are of considerable economic, social and environmental importance. Salmonids contribute to local and global economies through aquaculture, wild stock fisheries and recreational sport fisheries. In addition, they are a traditional food source for aboriginal peoples in Canada and have a central role in their culture. Salmon and trout are sentinel species for monitoring the aquatic environment and therefore they are used extensively for ecotoxicology studies. As a result of human activities related to the rearing of salmon and trout and the need to make management decisions concerning stock assessment and harvesting plans, there is a large salmonid research community working on the biology, life histories, population dynamics, biogeography, phylogenetic relationships, physiology and nutrition of salmonids.

Some fundamental scientific questions can be explored using salmonid genomes. The common ancestor of salmon and trout experienced a whole genome duplication, and modern species may be considered pseudo-tetraploid as they are in the process of reverting to a stable diploid state [2]. This makes them ideal organisms for examining the consequences of genome and gene duplications, processes that are considered to have had pivotal roles in generating gene diversity and the functional specialization found in modern vertebrates [3]. How a genome reorganizes itself to cope with duplicated chromosomes and the importance of gene duplications for evolution and adaptation are long-standing issues in biology that remain unresolved [3,4].

As illustrated in Table S1 in Additional file 1, no other group of fish species receives such comprehensive combined commercial and scientific attention as the salmonids [5], but as yet there is no genome sequence available for any salmonid. The genome of the Atlantic salmon (Salmo salar) was selected to be the reference sequence for all salmonids on the basis of its importance for the aquaculture industry and because so much previous work has been carried out at the genomic level on this species (Salmonid Genome Sequencing Workshop 2007, Simon Fraser University, Burnaby, BC). Having a robust salmonid reference sequence will allow next generation sequencing technologies to be used to obtain the sequences of other salmonids, such as rainbow trout and the Pacific salmon, more rapidly and at lower cost. An International Collaboration to Sequence the Atlantic Salmon Genome (ICSASG), representing researchers, funding agencies and industry from Canada, Chile and Norway, was formed to undertake this project. As suggested by the Toronto Statement on the pre-publication release of genomic data [6], in this open letter we provide details concerning: the approach adopted to sequence the Atlantic salmon genome; the plans for quality control of raw data, sequence assemblies, the integration of other genomic resources and annotation; and the types of analyses that are of particular interest to the ICSASG. It should be noted that the Toronto Statement [6] recognizes the right of the ICSASG to prepare the first large-scale description of the Atlantic salmon sequencing dataset and this open letter does not constitute that first publication.

Position of salmonids in teleost phylogeny and their relevance for vertebrate genome evolution
Fish and mammals share a common ancestor that dates back to the origin of the vertebrate lineage. With the
completion of the human genome sequence and several other mammalian genomes, it became clear that there is a need to investigate and characterize non-mammalian genomes to gain a more complete understanding of the complexity of vertebrate genomes and how they have evolved [7]. There are ~25,000 species of fish, making them the most successful vertebrate group, and their diversity makes them useful model systems in different disciplines of biology. The number of teleost species far exceeds those of any other fish group, or any other vertebrate, and this has been attributed to a whole genome duplication before their radiation in the Cretaceous Period [8-11]. In terms of phylogeny the salmonids hold a unique phylogenetic position compared with the fish species whose genomes have been sequenced or are in the process of being sequenced as they belong to the Protacanthopterygii, the most primitive group of teleosts (Figure 1). Thus, the salmonids provide a key phylogenetic link between teleost evolution and the evolution of non-teleost fish as well as other vertebrates.

**Properties of the Atlantic salmon genome and genomic resources currently available**

Much of the basic information concerning the Atlantic salmon genome is known. For example, the C value for Atlantic salmon has been estimated as 3.27 pg [12], which translates into a haploid genome size of $\sim 3 \times 10^9$ bp. The G+C content of the Atlantic salmon genome is 44.4% [13]. Although the Atlantic salmon genome is fairly similar to those of warm-blooded vertebrates with respect to size and overall base composition, it seems to be devoid of isochore structures like other coldwater fish genomes [14]. This is reflected in the inability to obtain
G-banding patterns in Atlantic salmon chromosomes [15]. It has been suggested that the diploid ancestor of salmonids possessed a karyotype with 48 acrocentric chromosomes, resulting in 96 acrocentrics after the genome duplication [16]. Comparisons of the karyotypes of several salmonid species, including Atlantic salmon, revealed that many gross chromosomal rearrangements (fusions and inversions) have occurred along the different lineages since the ancestral whole genome duplication occurred [16]. The Atlantic salmon whose genome has been chosen to be sequenced represents the European subspecies (S. salar europensis), with 29 pairs of chromosomes [17,18].

Approximately 200 cDNA libraries have been constructed from many different tissues and developmental stages of Atlantic salmon [19-27]. As of 2 April 2010, there were 495,257 Atlantic salmon Expressed Sequence Tags (ESTs) [28]. Atlantic salmon ranked 20th by organism with respect to total number of ESTs, and almost all of the other organisms in the top 20 have had their genomes sequenced or are in progress of so doing. The ESTs have been placed in over 81,000 contigs and annotated [25]. As of 2 April 2010, 33,709 Atlantic salmon UniGenes had been identified [29] and over 9,057 reference quality full-length cDNA coding sequences had been confirmed [27]. All of this information can be viewed on publicly available websites [29-32]. The EST databases provide a rich source of material for identifying genetic markers, such as microsatellites [33,34] and single nucleotide polymorphisms (SNPs) [35,36], that have been used to place genes on linkage maps [37,38]. In addition, the ESTs enabled the construction of microarrays for expression analyses [22,25,39-41]. More than 60 groups around the world are using these microarrays, indicating that there is a large salmonid research community actively engaged in functional genomics. Moreover, the EST databases, especially the full-length cDNA coding sequences, will form the basis for building gene models during the annotation of the Atlantic salmon genome.

A publicly available Atlantic salmon bacterial artificial chromosome (BAC) library (CHORI-214) was constructed from the DNA of an individual male from a Norwegian aquaculture strain and arrayed on nylon membranes [42]. HindIII fingerprinting of the CHORI-214 Atlantic salmon BACs was used to create the first physical map of a salmonid genome, consisting of 223,781 BACs in ~4,565 contigs and 33,217 singletons [43]. 207,869 BAC end sequences with an average length of 666 bp (~3.5% of the genome) were produced to yield a snapshot of the Atlantic salmon genome and to identify putative syntenic relationships between the Atlantic salmon physical map and the fish genomes that have been sequenced ([44] and KA Boroevich, KP Lubieniecki, W Chow, P de Jong, J Schein, M Field, R Moore, JG de Boer, BFK, WSD, unpublished results). Several linkage maps based on microsatellites, amplified fragment length polymorphisms (AFLPs) and SNPs have been constructed for Atlantic salmon [37,38,45-47]. The BAC end sequences also provide a rich source of microsatellite markers and SNPs, which were used to integrate the physical and linkage maps [37,47]. Fluorescent in situ hybridization (FISH) analysis with BACs that contain microsatellite markers map was used to assign chromosome arms to linkage groups [48]. All of these genomic resources have been made publicly accessible through a website ([49] and KA Boroevich, KP Lubieniecki, W Chow, P de Jong, J Schein, M Field, R Moore, JG de Boer, BFK, WSD, unpublished results).

An Illumina iSelect bead-array, designed to interrogate ~16,500 putative Atlantic Salmon SNPs, was developed at the Center for Integrative Genetics (CIGENE). Approximately 55% of the SNPs on the array were identified from EST alignments, with most of the remainder coming from a random genomic sampling following construction of reduced representation libraries produced from individual and pooled DNA samples and high-throughput 454 pyrosequencing (MP Kent, B Hayes, Q Xiang, PR Berg, RA Gibbs, S Lien, personal communication). The SNP array is currently being used to genotype a mapping population consisting of 3,500 individuals and construct a high-resolution SNP map for Atlantic salmon, which will be beneficial in assembling the Atlantic salmon genome sequence.

More than 60 Atlantic salmon BACs have been sequenced so far ([50-58], and see [59] for list). Along with the BAC end sequences these provide a snapshot of the organization of the Atlantic salmon genome. It is estimated that repetitive DNA accounts for 30-35% of the Atlantic salmon genome. Fourteen families of DNA transposons (twelve TcI-like and two piggyBac-like) constituting 6-10% of the genome have been identified. These DNA transposons are approximately 1,500 bp in length, with different repeat families ranging in similarity from 80% to 94% [60]. An Atlantic salmon repeat database has been developed, and a salmonid repeat masking tool is publicly available [61].

The Atlantic salmon genome sequencing strategy

Although five fish genomes have been sequenced, they represent euteleostei lineages that have been separated from salmonids for at least 200 million years (Figure 1) [62]. The C values of these fish genomes range in size from 0.35 pg (spotted green pufferfish) to 1.80 pg (zebrafish) [63], whereas the Atlantic salmon genome is considerably larger (3.27 pg) [12]. The Atlantic salmon genome is further complicated by the autotetraploid whole genome duplication, which occurred ~25-100 million years ago in the common ancestor of extant
salmonids [2], and the ongoing re-diploidization process, which involves genome rearrangements and the loss, sub-functionalization and neo-functionalization of duplicated genes [3,4,64]. Comparisons of duplicated regions of the Atlantic salmon genome reveal that they are 81-89% identical (e.g. 81-85% identity over 225 kb in the IgH A and B regions [56] and 87-89% identity in the duplicate copies of growth hormone genes [52]). Similarly, EST contig duplicates show 91-93% identity [25,27]. Therefore, it should be possible to resolve and assemble complex duplicated regions of the Atlantic salmon genome. Nevertheless, the size and complexity of the Atlantic salmon genome, particularly the long and frequent repeats [60], combined with the lack of a closely related guide sequence, means that assembling the sequence of the Atlantic salmon genome will be challenging.

Some of the problems of assembling and characterizing duplicate segments of the genome will be overcome by the choice of the Atlantic salmon to be sequenced: a double-haploid individual that was produced by mitotic androgenesis. The fish was female and was nicknamed ‘Sally’. The homozygous nature of Sally’s genome was verified by screening for polymorphisms at ~70 micro-satellite loci. Karyotyping of Sally revealed that she had a haploid chromosome content of 29, in accordance of what would be expected for an Atlantic salmon from Norway. No apparent chromosomal rearrangements were observed; however, she does seem to have been a mosaic with ~30% haploid cells and the remainder diploid cells (U Grimholt, personal communication). The availability of DNA from a totally homozygous individual will greatly facilitate the assembly of whole genome shotgun sequences. Because the fish chosen for sequencing was female and male salmon are the heterogametic sex [65], the male sex-determining region will not be sequenced at this stage of the Atlantic salmon genome sequencing project.

The feasibility of using GS FLX pyrosequencing (shotgun and paired-end reads with an average read length of 250 bp) to sequence the Atlantic salmon genome was assessed by attempting to sequence six BACs that form a minimum tiling path estimated to be ~1 Mb [66]. The conclusion reached was that the GS FLX technology was limited to gene mining and establishing a set of ordered sequence contigs with many gaps. Although the pyrosequencing technology has been improved and average read lengths of 450 bp are now routinely achieved, this is still not sufficient to get at least halfway through a repeat of 1,500 bp, especially if paired end reads are desired. Indeed, the recent assembly of the cod genome seems to confirm this. A 27-fold coverage based on GS FLX Titanium reads (average read length of 336 bp), which included paired ends, produced contigs with an N₅₀ of 2,400 bp and ~14,000 scaffolds with an N₅₀ of 571 kb that cover 618 Mb of the predicted 930 Mb genome (where N₅₀ is defined as the contig length such that using equal or longer contigs produces half the bases of the genome). The N₅₀ size is computed by sorting all contigs from largest to smallest and by determining the minimum set of contigs whose sizes total 50% of the entire genome. Currently, to obtain a genome sequence that can act as a reference sequence for other salmonids, it seemed that a substantial portion of the sequencing of the Atlantic salmon genome should be carried out using Sanger technology or an equivalent with respect to read length. The repetitive nature of the Atlantic salmon genome, and especially the length of the most common repeat (~1,500 bp) [60], make it necessary to have long paired end reads for assembling the sequence of this species’ genome.

Description of sequencing project with anticipated milestones
The ICSASG has raised sufficient funds from the Research Council of Norway, the Norwegian Fishery and Aquaculture Industry Research Fund, Genome British Columbia, the Chilean Economic Development Agency and the InnovaChile Committee to cover the cost of sequencing, assembling and annotating the Atlantic salmon genome. In Phase 1, which began on 1 January 2010, Beckman Coulter Inc. will produce a fourfold coverage of the genome using paired end, plasmid, fosmid and BAC Sanger sequences by the end of January 2011. Phase 2 will be conducted using primarily next generation sequencing technologies and will result in a high definition, well-annotated genome. It is anticipated that Phase 2 will commence in January 2011 and that the sequencing component will be completed by September 2011. The assembled sequence will be integrated with the physical map, the linkage map and the ESTs. By the beginning of 2012 an automated pipeline annotation will be prepared and the sequence placed on a genome browser such as Ensembl [67]. The salmonid community will then be invited to participate in a manual annotation of the Atlantic salmon genome using the strategy advocated by Elsik et al. [68]. We anticipate that the key paper and companion reports describing the Atlantic salmon genome and its insight into salmonid biology and vertebrate genome evolution will be published in the summer of 2012.

Biological questions and types of analyses to be undertaken by ICSASG
The availability of a complete genome sequence for Atlantic salmon will have a major impact on all sectors of the international salmonid community. For the aquaculture industry it will provide a complete suite of genetic markers for the identification of the genes and alleles
responsible for production traits (such as growth, disease resistance, feed efficiency and age of sexual maturation). Companies will be able to develop tailored broodstock more precisely using nucleotide- or allele-assisted selection rather than the more general marker-assisted selection. In conjunction with traditional breeding practices, this approach promises rapid gains that will make companies who embrace this technology more competitive. For those government agencies with a mandate to conserve and manage wild stocks, the sequence will provide the tools that will make it possible to identify and distinguish discrete populations of salmon using genes that are selected for specific environments (such as thermal tolerance) rather than neutral genetic markers. Thus, the genome sequence has great potential for enabling sensitive and more accurate assessments of the sustainability of salmonid populations. For the academic community, the Atlantic salmon genome will provide a reference sequence for other related genomes (such as rainbow trout, Pacific salmon or char), which could be rapidly and cost effectively sequenced using novel sequencing technologies. This will allow comparative genomics to be incorporated into ecology and evolutionary and population biology and bring to the fore the concept of landscape evolutionary genomics [69]. For ecotoxicologists, the salmon sequence will permit a more robust use of salmonids as sentinel species for monitoring the quality of the aquatic environment using genes whose expression patterns are known to respond to particular pollutants.

Some specific questions and projects that researchers associated with the ICSASG are particularly interested in, and which provided the motivation for sequencing the Atlantic salmon genome, include: the characterization of the immune system in salmon [50,51,56-58] and how this relates to resistance to specific pathogens [44,70,71] that affect the aquaculture industry; the identification of the master sex-determining gene and the pathways that regulate sexual maturation [65,72-76]; an understanding of the biological cues and the sensing mechanisms that allow salmon to return faithfully to their natal streams after extensive marine migrations [53,54]; and the rediploidization process and the fates of duplicated genes after a whole genome duplication [25,27,52,55].

**Formation of the ICSASG and a portal for salmonid genomic resources**

Individual researchers were receiving grants from national agencies during the 1980s and 1990s to carry out genetic studies on salmonids, but the funding was not sufficient for any group to make significant headway on their own. The Norwegian Salmon Genome Project and the Canadian Genomics Research on Atlantic Salmon Project (GRASP), funded from 2001 until 2005, allowed great progress to be made, but it was evident to the participants that even more could be achieved if they pooled resources and worked together. This was the basis for forming cGRASP: the Consortium for Genomic Research on All Salmonids Project, which was successful in bringing together salmonid genomics teams from Canada, Norway, Scotland and the USA.

In 2005 the need for a scientific organizational body to coordinate genome research efforts and ensure that existing and upcoming resources were made accessible worldwide was recognized by the international salmonid research community. At a workshop held on 25-26 October 2005 at the Norwegian University for Life Sciences in Ås, Norway, the Consortium for Genomic Research on All Salmonids Program (cGRASP) was formed. A follow-up meeting held on October 10-12, 2006 at Simon Fraser University, Burnaby, Canada, attracted representatives of the salmonid research community from 17 countries. These researchers identified very clearly that there was a need for at least one high quality, whole genome salmonid sequence to make optimal use of genomics tools within salmonid research, that many of the pre-sequencing phase resources would be in place for Atlantic salmon by the end of 2008 and that the research community was strongly committed to building and maintaining the necessary organizational apparatus for handling the pre-sequencing phase, the draft sequence phase and the initial annotation phase.

Given the benefit that the salmon genome sequence will bring to the aquaculture industry, government agencies charged with managing wild stocks and monitoring the aquatic environment, foundations concerned with conservation issues and the academic community at large, it was proposed that an international public-private partnership would be the most appropriate funding model to accomplish this task. At a meeting in Quebec City in 2008 representatives of funding agencies from British Columbia, Norway and Chile resolved to work together to sequence the Atlantic salmon genome, and in April 2009 in Santiago, Chile the ICSASG was formally established. The ICSASG has raised sufficient funds to cover the cost of sequencing, assembling and annotating the Atlantic salmon genome. It is anticipated that when the sequence becomes available, other opportunities will arise and the framework established by the ICSASG can be expanded to encompass projects such as sequencing other salmonid genomes using the Atlantic salmon as a reference sequence. After the meeting at Simon Fraser University in 2006 a website [77] was set up as a portal for other websites that host salmonid genomic data, as well as providing information concerning on-going projects, collaborative opportunities and contact information for the ICSASG.
Salmolo salar L. project (GRASP) microarray tools. J. Fish Biol 2008, 72:2051-2070.

41. Taggart JB, Bron JE, Martin SA, Seear PJ, Hayeheim B, Talbot R, Carmichael SN, Villeneuve LA, Sweeney GE, Houlihan DF, Seecombe CJ, Tocher DR, Teale AJ. A description of the origins, design and performance of the TRAITS-SGP Atlantic salmon L. DNA microarray. J. Fish Biol 2008, 72:2071-2094.

42. Thorsen J, Zhu B, Frenzen E, Osogewa K, de Jong Pi, Koop BF, Davidson WS, Hoyeheim B. A highly redundant BAC library of Atlantic salmon (Salmolo salar): an important tool for salmon projects. BMC Genomics 2005, 6:50.

43. Ng SHS. Morphometric, histological and cytochemical characterization of the sex determining loci and sex chromosomes in the family salmonidae. J Mol Evol 2009, 68:118-125.

44. Lai YYY, Lubieniecki KP, Chow W, Koop BF. Assignment of Atlantic salmon (Salmolo salar) linkage groups to specific chromosomes: conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (Onchorhynchus mykiss). BMC Genet 2009, 10:46.

45. Asalbase [http://www.asalbase.org].

46. Yazawa R, Cooper GA, Hunt P, Beetz-Sargent M, Robb A, Conrad M, McKinnel L, So S, Jantzen S, Phillips RB, Koop BF. Striking antigen recognition diversity in the Atlantic salmon T-cell receptor alpha/delta locus. Dev Comp Immunol. 2008, 32:204-212.

47. Artieri CG, Bosdet IE, Chiu R, Danzmann RG, Davidson WS, Ferguson MM, Fjell CD, Hoyeheim B, Clarke L, Loosli JO, Halko S, Simon M, Myllyvirta S, Suomalainen R, Shahin M, Weichelt U, Heino M, Peltola T, Brandbyge M, Kasprzyk A, Kahari A, Saarinen J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari V, Keefe D, Keenan S, Lehmavaho H, McVicker G, Melsopp C, Meidi P, Morgan E, Pettett R, et al. An overview of Ensembl. Genome Res 2004, 14:925-928.

48. Phillips RB, Keatley KA, Morasch MR, Ventura AB, Lubieniecki KP, Koop BF, Fjell CD, Hoyheim B, Jones SJM, de Jong PJ, Phillips BB, Garrett JA, Lien L, Jantzen S, Phillips RB, Davidson WS, Koop BF. Asalbase (http://www.asalbase.org). Atlantic salmon L. project (GRASP) microarray tools. J Mol Evol 2008, 67:117-125.

49. Lai YYY, Lubieniecki KP, Phillips RB, Chow W, Koop BF. Atlantic salmon L. genome and evolution of the vromeronasal type 2 receptor-like (OVIC) gene clusters in Atlantic salmon. Mol Biol Evol 2009, 26:1177-1187.

50. Artieri CG, Mitchell LA, Ng SHS, Pariosottto SE, Danzmann RG, Hoyeheim B, Phillips RB, Morasch M, Koop BF. Identification of the sex-determining locus of Atlantic salmon (Salmolo salar). J Mol Evol 2009, 69:1221-1229.

51. Yazawa R, Cooper GA, Beetz-Sargent M, Robb A, McKinnel L, Davidson WS, Koop BF. Striking antigen recognition diversity in the Atlantic salmon T-cell receptor alpha/delta locus. Dev Comp Immunol. 2008, 32:204-212.

52. Johnstone KA, Lubieniecki KP, Chow W, Phillips RB, Koop BF, Davidson WS. Genetic origin and characterization of two vromeronal type 1 receptor-like genes (orol1 and orol2) in Atlantic salmon. Mol Genet Genomics 2008, 279:1231-1237.