

An Integrated Approach to Gene Discovery and Marker Development in Atlantic Cod (*Gadus morhua*)

Sharen Bowman · Sophie Hubert · Brent Higgins · Cynthia Stone · Jennifer Kimball · Tudor Borza · Jillian Tarrant Bussey · Gary Simpson · Catherine Kozera · Bruce A. Curtis · Jennifer R. Hall · Tiago S. Hori · Charles Y. Feng · Marlies Rise · Marije Booman · A. Kurt Gamperl · Edward Trippel · Jane Symonds · Stewart C. Johnson · Matthew L. Rise

Received: 4 January 2010 / Accepted: 4 March 2010 / Published online: 16 April 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract Atlantic cod is a species that has been over-exploited by the capture fishery. Programs to domesticate this species are underway in several countries, including Canada, to provide an alternative route for production. Selective breeding programs have been successfully applied in the domestication of other species, with genomics-based approaches used to augment conventional methods of

animal production in recent years. Genomics tools, such as gene sequences and sets of variable markers, also have the potential to enhance and accelerate selective breeding programs in aquaculture, and to provide better monitoring tools to ensure that wild cod populations are well managed. We describe the generation of significant genomics resources for Atlantic cod through an integrated genomics/

Electronic supplementary material The online version of this article (doi:10.1007/s10126-010-9285-z) contains supplementary material, which is available to authorized users.

S. Bowman (✉) · S. Hubert · B. Higgins · C. Stone · T. Borza · J. T. Bussey · G. Simpson · C. Kozera · B. A. Curtis
The Atlantic Genome Centre,
NRC Institute for Marine Biosciences, 1411 Oxford Street,
Halifax, NS B3H 3Z1, Canada
e-mail: sbowman@genomeatlantic.ca

S. Hubert
Department des Sciences, Université Sainte Anne,
Pointe-de-l'Église, NS B0W 1M0, Canada

J. Kimball
The Atlantic Genome Centre,
NRC Institute for Marine Biosciences,
National Research Council,
1411 Oxford Street,
Halifax, NS B3H 3Z1, Canada

J. R. Hall · T. S. Hori · C. Y. Feng · M. Rise · M. Booman · A. K. Gamperl · M. L. Rise
Ocean Sciences Centre, Memorial University of Newfoundland,
1 Marine Lab Road,
St. John's, NL A1C 5S7, Canada

E. Trippel
Fisheries and Oceans Canada,
St. Andrews Biological Station, 531 Brandy Cove Road,
St. Andrews, NB E5B 2L9, Canada

J. Symonds
National Institute of Water and Atmospheric Research,
NIWA Bream Bay Aquaculture Park, PO Box 147,
Ruakaka 0151, New Zealand

S. C. Johnson
Pacific Biological Station, Fisheries and Oceans Canada,
Nanaimo, BC V9T 6N7, Canada

Present Address:
J. T. Bussey · C. Kozera
The Atlantic Genome Centre,
NRC Institute for Marine Biosciences,
National Research Council,
1411 Oxford Street,
Halifax, NS B3H 3Z1, Canada

Present Address:
B. A. Curtis
Department of Biochemistry and Molecular Biology,
Dalhousie University,
5850 College Street,
Halifax, NS B3H 1X5, Canada

selective breeding approach. These include 158,877 expressed sequence tags (ESTs), a set of annotated putative transcripts and several thousand single nucleotide polymorphism markers that were developed from, and have been shown to be highly variable in, fish enrolled in two selective breeding programs. Our EST collection was generated from various tissues and life cycle stages. In some cases, tissues from which libraries were generated were isolated from fish exposed to stressors, including elevated temperature, or antigen stimulation (bacterial and viral) to enrich for transcripts that are involved in these response pathways. The genomics resources described here support the developing aquaculture industry, enabling the application of molecular markers within selective breeding programs. Marker sets should also find widespread application in fisheries management.

Keywords Atlantic cod · ESTs · Markers · SNPs · Aquaculture

Introduction

The depletion of Atlantic cod stocks in the North Atlantic has resulted in reduced harvests of this species by the capture fishery (Rose 2007; Myers et al. 1997), thus providing an incentive to develop commercial farming for Atlantic cod. Initiatives to domesticate this species are underway in several countries including Canada and Norway (Rosenlund and Skretting 2006; Brown et al. 2003) using breeding to select for commercially desirable phenotypes such as high food conversion efficiencies, increased stress tolerance and enhanced disease resistance. The aquaculture community needs to overcome several challenges before a commercially viable industry is achieved for Atlantic cod. Infectious agents causing significant losses in production of this species in captivity include nodavirus (Patel et al. 2007), *Aeromonas salmonicida* subspecies (including atypical strains), and *Vibrio* species (Rodger and Colquhoun 2008), with emerging pathogens frequently being identified, such as the microsporidian *Loma* sp. (Khan 2005; Rodriguez-Tovar et al. 2003). Atlantic cod can also experience variations in temperature at mariculture sites. Fish confined within sea cages cannot avoid these non-optimal temperatures and thus experience stress (Gollock et al. 2006). Another stressor of fish grown in captivity is handling stress, which is encountered through routine aquaculture practices such as fish transfers, grading and tagging (Brown et al. 2003). In addition, issues such as early maturation can have a significant effect on production, as spawning fish undergo considerable weight reduction and loss of condition (Hansen et al. 2001).

Genomics tools have the potential to enhance and accelerate selective breeding programs through the identi-

fication of quantitative trait loci (QTL) and the incorporation of marker assisted selection (MAS). Many expressed sequence tag (EST) sequencing programs in commercially important species have already been completed (Rise et al. 2004; Gorodkin et al. 2007; Park et al. 2005; Koop et al. 2008). These sequence collections are regularly mined for marker development (He et al. 2003; Snelling et al. 2005; Hayes et al. 2007; Ramsey et al. 2007; Moen et al. 2008) with the aim of using these markers to enhance desirable characteristics through MAS. However, often sequencing initiatives have not been designed to maximise the opportunities for marker identification within breeding programs; frequently ESTs have been generated from varieties or individuals which have not been included in mapping populations or breeding programs. This necessitates an additional phase of directed re-sequencing to identify single nucleotide polymorphisms (SNPs) or microsatellite length variants that are present in the target individuals or populations (Labate and Baldo 2005; Lijavetzky et al. 2007).

We have attempted to address this issue by initiating an integrated genomics (Bowman et al. 2007) and broodstock development program (Symonds et al. 2007), the Atlantic Cod Genomics and Broodstock Development Project (CGP), with the generation of genomics resources designed to be applied directly in two family-based breeding programs located in New Brunswick (NB) and Newfoundland (NL), Canada. The rationale for undertaking separate province-based breeding programs has its basis in the observation of genetic differences among cod stocks in the Northwest Atlantic (Hutchinson et al. 2001; Ruzzante et al. 1996). This has resulted in a desire to minimise the potential for interbreeding of resident wild stocks with captive stock (either as a result of escapees, or via the release of eggs or sperm from fish in sea cages) that might be ill adapted to the local environmental conditions by using local stocks for breeding. ESTs have been generated from fish enrolled in the breeding programs, allowing identification of genetic markers present in family fish as a direct outcome from the sequencing component of the project, enabling genetic mapping, analysis of QTL and ultimately MAS.

This study describes the construction of cDNA libraries and the generation of a large EST set for Atlantic cod. The majority of the sequencing effort focused on normalized libraries to maximise gene identification (Bonaldo et al. 1996; Zhulidov et al. 2004), with sequences generated from multiple tissues and developmental stages. Several libraries were produced from tissues sampled after exposure to stressors or antigen stimulation to ensure that transcripts important in aquaculture-relevant traits such as disease resistance and temperature susceptibility/tolerance were represented in our collection. Several suppression subtractive hybridisation (SSH) libraries (Diatchenko et al. 1996) were also generated for targeted gene discovery i.e. for

identification of genes that are induced or suppressed in physiologically relevant tissues following exposure to a viral pathogen or the viral mimic polyribonucleosinic polyribocytidylic acid (pIC) (Rise et al. 2008), bacterial antigens (Feng et al. 2009), or heat-shock (Hori et al. 2010). A small number of these sequences, notably several of the SSH libraries, have been analysed separately (Rise et al. 2008; Feng et al. 2009; Hori et al. 2010), however approximately 130,000 of the ESTs presented here have not been described previously, and this is the first overall analysis of the entire data set. A set of novel microsatellite markers has already been identified from this dataset (Higgins et al. 2009). Putative SNP markers identified in the sequence data were tested by genotyping and found to be variable in Canadian and North European populations and in fish enrolled within the two family-based breeding programs from which the sequence resource was developed.

Materials and Methods

Tissue Sampling

Fish for sampling were held at the St. Andrews Biological Station, St. Andrews, New Brunswick, Canada, at the Ocean Sciences Centre, Memorial University, Newfoundland, Canada, or at Great Bay Aquaculture, Newington, New Hampshire, USA. Tissue samples (approximately 0.5 cm³) were taken from fish that had been euthanized using a lethal dose of anaesthetic. Samples were incubated in RNAlater (Ambion Inc., Austin, TX) overnight at 4°C, or flash frozen in liquid nitrogen, prior to storage at -80°C. The heat shock and immunogenic challenge procedures used in the production of a subset of libraries have been described in detail in Rise et al. 2008, Feng et al. 2009 and Hori et al. 2010.

Experimental manipulations and sampling of the fish were carried out in accordance with the Canadian Council on Animal Care Guidelines and approved by animal care committees at Dalhousie University, The Institute of Marine Biosciences, Halifax, and Memorial University of Newfoundland.

Production of Normalized cDNA and SSH Libraries

Tissue samples from multiple fish (80–100 mg individual⁻¹) were pooled prior to RNA extraction in the construction of normalized libraries. Total RNA was extracted from tissue pools by lysis using TRIzol reagent (Invitrogen, Burlington, ON) followed by chloroform extraction and precipitation with isopropanol as specified in the manufacturer's protocol. Normalized cDNA library generation was performed essentially as described in Bowman et al. 2007. cDNA production and library construction were carried out

using the Creator SMART kit (Clontech, Mountain View, CA). cDNA was normalized using the Trimmer kit method (Evrogen JSC, Moscow, Russia) (Zhulidov et al. 2004) prior to directionally cloning into the SfiI sites of pDNR-LIB (Clontech). Methods utilized to construct reciprocal SSH libraries, including total RNA isolation, mRNA isolation, suppression subtractive hybridization, cloning, and initial evaluation of SSH library insert size and complexity, have been previously published (Rise et al. 2008; Feng et al. 2009; Diatchenko et al. 1996; Hori et al. 2010).

Normalized library names were generated using a series of two letter codes to identify production variables. For example: gmnbhkic is identified as a library created from *Gadus morhua* (gm) fish, from the New Brunswick Breeding Program (nb), head kidney (hk) was the tissue used, and the fish had undergone a treatment with pIC (ic). The SSH libraries follow similar nomenclature rules with respect to fish species, fish breeding program, and fish treatment; however, in addition, an “f” or an “r” was used in this naming convention to denote the “forward” or “reverse” subtracted libraries (i.e. enriched for transcripts that were up-regulated or down-regulated by a stressor, respectively), and only one letter was used to identify tissue type instead of two. For example: gmnlfkic is identified as a SSH library created from *G. morhua* (gm) fish, from the Newfoundland Breeding Program (nl), head kidney (k) was the tissue used, it was a forward SSH library (f), and the fish had undergone a treatment with pIC (ic).

DNA Sequencing

DNA sequencing was performed as described in Bowman et al. 2007. Briefly, individual bacterial clones were inoculated into 384 well plates containing LB/glycerol, plus an appropriate antibiotic to select for growth of transformed clones, using a QPIX robotic clone picker. After incubation overnight, clones were stored at -80°C prior to sequencing. DNA was prepared from each clone using templiphi (GE Healthcare, Piscataway, NJ), with sequencing reactions performed using ET terminator chemistry (GE Healthcare). After purification to remove excess fluorescent terminator reagents, samples were separated by electrophoresis on MegaBACE capillary sequencers (GE Healthcare).

EST Clustering and Annotation

Sequence trace files were basecalled using Phred (Ewing et al. 1998; Ewing and Green 1998), with poor quality sequences trimmed using Phred and vector sequences masked using Paracel Transcript Assembler (PTA). The resulting ESTs were clustered using PTA. Initially, normalized libraries were sequenced from the 3'

end. Clustering of these 3' sequences was performed on a per-library or per-tissue basis with an additional clustering run of all 3' ESTs. The contigs and singletons generated from clustering of 3' sequences from all ESTs were used to select a set of unique clones for additional 5' sequencing. This involved the clones for all singletons being selected for 5' sequencing with, in addition, one clone from each contig also being selected (the clone representing the extreme 5' end of that contig). Selected clones were rearranged using a QPIX robot to inoculate wells from the original plates into new 384 well plates containing LB/glycerol/antibiotic. Sequencing and trace processing were carried out as described earlier. Additional clustering runs of 3' and 5' sequences were then performed for each normalized library, and a clustering run of all EST data (3', 5' and SSH) was also generated. ESTs have been deposited in the dbEST subsection of GenBank under accession numbers shown in Supplemental Table S1.

Sequences and contig consensi were processed through an automated annotation pipeline based on AutoFACT (Koski et al. 2005). This program processes the output of multiple database searches and generates a consensus annotation. Databases used in AutoFACT included UniRef90 (Uniprot), nr (NCBI), Kyoto Encyclopedia of Genes and Genomes (KEGG), Clusters of Orthologous Groups (COG), PFAM, LSU and SSU. BLAST hits were classified as significant in AutoFACT if they had a bit score of 40 or higher. Annotation of the complete dataset can be accessed via the project database Codgene (<http://ri.imb.nrc.ca/codgene>). Codgene displays information on all libraries constructed, and includes annotation summaries, access to relevant sequences by keyword searching, BLAST searching of selected datasets and access to statistics generated for libraries, sequences and clustering runs.

SNP Identification and Genotyping

SNPs were identified from clusters of four or more 3' sequences with a predicted minor allele frequency greater than 25% as described in Hubert et al. 2009. Two GoldenGate panels (Illumina, San Diego, CA) of 1536 predicted SNPs were constructed, with high-throughput genotyping carried out at the Genome Quebec and McGill University Innovation Centre. DNA for genotyping was prepared from fin clips taken from fish collected around the Atlantic Canadian coast, including Cape Sable (NB YC1), Georges Bank (NB YC2), Bay Bulls (NL YC2) and Smith Sound (NL YC3), and from Northern Europe, including Iceland, Ireland and Norway using the Qiagen DNeasy tissue kit (Mississauga, ON) according to the manufacturer's instructions.

Results

EST Sequencing

In total, we generated 158,877 sequences for Atlantic cod (see Supplementary Table S1 for a complete set of accession numbers). Of these, 138,760 were generated from 23 normalized libraries, and 20,117 from 19 SSH libraries (Table 1). These libraries were generated from 12 tissue types, and four developmental stages, including embryos, larvae, juveniles and adults. To maximise identification of genes relevant to desirable traits for aquaculture, some samples were taken from fish subjected to stressors (e.g. net or heat shock) (Hori et al. 2010) or exposures to immunogenic stimuli (e.g. pIC (Rise et al. 2008) or formalin-killed, atypical *A. salmonicida* (Feng et al. 2009)). Several SSH-identified transcripts have been shown by quantitative reverse transcription polymerase chain reaction to be responsive to immunogenic stimuli (Feng et al. 2009; Rise et al. 2008) or heat shock (Hori et al. 2010), indicating that our targeted gene discovery approach was effective for the identification of genes important in the cellular responses to such stimuli.

To facilitate marker development (e.g. SNP discovery), normalized cDNA libraries were generated from tissue samples isolated from numerous fish so that sequence variation between individuals could be identified. The number of individuals from which tissue was pooled prior to RNA extraction ranged from three to 340, with an average of 41. RNA used for SSH library generation was pooled after extraction. Individuals sampled were taken either from the same populations as the fish used for the production of families in NB and NL, or from fish enrolled in the family programs (for example parents of families or F1 offspring) to ensure that the markers identified were present in the family programs. New Brunswick parent fish contributing to the sequence data were collected from Northwest Atlantic Fisheries Organisation (NAFO) subdivision 4X (Myers et al. 1997) (Cape Sable in south west Nova Scotia) and were used to generate NB year class 1 families in 2006 (NB YC1). Newfoundland fish contributing to the sequence data were collected from NAFO subdivision 3L (Bay Bulls, NL) and were used the following year to generate NL YC2 families (Tables 1, 2 and 3).

Normalized libraries were initially sequenced from the 3' direction, which facilitated clustering and ensured access to the 3' untranslated region (3'UTR), where we anticipated finding a lower level of splicing (Kan et al. 2000) and a higher level of polymorphism for the purpose of marker identification. The disadvantages of this approach include reduced sequence quality, more sample failures, shorter read lengths and fewer informative hits to sequence data-

Table 1 List of CGP libraries

Library Name	Library Type	Breeding Program	Tissue Type	For./Rev.	Tissue Treatment	No. of sequences
gmnsil	N	N/A	noda. infected liver	N/A	none	381
gmnsul	N	N/A	uninfected liver	N/A	none	136
gmnbbr	N	NB	brain	N/A	none	9322
gmnbgi	N	NB	gill	N/A	none	2224
gmapht	N	NB,NL,NH	heart	N/A	none	16983
gmapov	N	NB,NL,NH	ovary	N/A	none	11555
gmnbbrts	N	NB	brain	N/A	thermal stress	3358
gmnbllts	N	NB	liver	N/A	thermal stress	15475
gmnbhkas	N	NB	head kidney	N/A	<i>A. sal.</i>	6057
gmnbhkie	N	NB	head kidney	N/A	pIC	7510
gmnbmd	N	NB	mixed digestive	N/A	none	6151
gmnbmu	N	NB	muscle	N/A	none	4600
gmapte	N	NB,NL,NH	testis	N/A	none	4225
gmnlkic	N	NL	spleen/head kidney	N/A	pIC	4665
gmlbgits	N	NB,NL	gill	N/A	thermal stress	1168
gmnbpcic	N	NB	pyloric caeca	N/A	pIC	4561
gmnbpic	N	NB	spleen	N/A	pIC	4544
gmnl2pia	N	NL	blood	N/A	pIC & <i>A. sal.</i>	2538
gmnlm	N	NL	embryo	N/A	none	12135
gmnlia	N	NL	larvae	N/A	none	20493
gmnlpbas	N	NL	blood	N/A	<i>A. sal.</i>	71
gmnlpbia	N	NL	blood	N/A	pIC & <i>A. sal.</i>	513
gmnlpbic	N	NL	blood	N/A	pIC	95
gmnlkfic	SSH	NL	head kidney	F	pIC	1005
gmnlkric	SSH	NL	head kidney	R	pIC	135
gmnlkfic	SSH	NL	spleen	F	pIC	3005
gmnlkric	SSH	NL	spleen	R	pIC	869
gmnlkfta	SSH	NL	head kidney	F	heat shock	1451
gmnlkrta	SSH	NL	head kidney	R	heat shock	93
gmnlifta	SSH	NL	liver	F	heat shock	1524
gmnlirta	SSH	NL	liver	R	heat shock	1586
gmnlmfta	SSH	NL	skeletal muscle	F	heat shock	1419
gmnlbfic	SSH	NL	brain	F	pIC	2067
gmnlbric	SSH	NL	brain	R	pIC	2111
gmnlkfas	SSH	NL	head kidney	F	<i>A. sal.</i>	1033
gmnlkras	SSH	NL	head kidney	R	<i>A. sal.</i>	986
gmnlpfas	SSH	NL	blood	F	<i>A. sal.</i>	81
gmnlpras	SSH	NL	blood	R	<i>A. sal.</i>	85
gmnlpfic	SSH	NL	blood	F	pIC	442
gmnlpric	SSH	NL	blood	R	pIC	90
gmnlkfas	SSH	NL	spleen	F	<i>A. sal.</i>	1048
gmnlkfas	SSH	NL	spleen	R	<i>A. sal.</i>	1087

For each library, the type, breeding program, tissue(s) used and any treatments used on fish prior to tissue collection are shown, together with the number of sequences generated. Libraries produced were generated using either normalized library (N) or suppression subtractive hybridization (SSH) protocols, with the latter being divided into forward (F) or reverse (R) libraries. Fish were collected for the New Brunswick (NB), Newfoundland (NL) or New Hampshire (NH) breeding programs, or originated from a pre-project collection (N/A). The scheme used for library nomenclature is described in the Methods section. N/A = not applicable, noda. = nodavirus. For details on the heat shock challenge, see Hori et al 2010. For details on the formalin-killed, atypical *Aeromonas salmonicida* (*A. sal.*) challenge, see Feng et al. 2009. For details on the viral mimic (pIC) challenge, see Rise et al. 2008.

Table 2 Normalized cDNA libraries created for *Gadus morhua* contributing sequences used for SNP identification

Library Name	No. of Fish	Breeding Program	Age	Fish Population	No. of sequences
gmnsil	3	N/A	Adult	Pre-project collection from NS waters	178
gmnsul	3	N/A	Adult	Pre-project collection from NS waters	67
gmnbbr	10	NB program	Adult	Same population as NB YC1 parents	5992
gmnbgi	10	NB program	Adult	Same population as NB YC1 parents	1346
gmapht	20	NB, NL, NH program	Adult	Same population as NB YC1 and NL YC2 parents	12374
gmapov	20	NB, NL, NH program	Adult	Same population as NB YC1 and NL YC2 parents	7683
gmnbbrts	14	NB program	Juvenile	NB YC1 F1 progeny	1959
gmnbblts	14	NB program	Juvenile	NB YC1 F1 progeny	10650
gmnbhkas	14	NB program	Juvenile	NB YC1 F1 progeny	4665
gmnbhkie	18	NB program	Juvenile	NB YC1 F1 progeny	5460
gmnbmd	10	NB program	Adult	NB YC1 parents	4148
gmnbmu	10	NB program	Adult	NB YC1 parents	3720
gmnbpic	18	NB program	Juvenile	NB YC1 F1 progeny	3833
gmnbpic	14	NB program	Juvenile	NB YC1 F1 progeny	3858
gmapte	17	NB, NL, NH program	Adult	Same population as NB YC1 and NL YC2 parents	3279
gmnlkie	21	NL program	Juvenile	NL YC1 F1 progeny	3795
gmnlm	340	NL program	Embryo	NL YC2 F1 progeny	8958
gmnlia	290	NL program	Larvae	NL YC2 F1 progeny	15550
gmnlbigs	12	NB, NL program	Juvenile	NB YC1 F1 progeny, NL YC1 F1 progeny	241
gmnlpia	26	NL program	Juvenile blood	NL YC1 F1 progeny	240

Libraries were generated by pooling one (or more) tissue types from multiple fish, from one (or more) breeding programs. YC1 and YC2 denote Year Class 1 and Year Class 2 fish, respectively.

bases. To allow better informative annotation of the sequence generated, a clone representing each singleton, or a single clone selected from the extreme 5' end of each unique contig, was also sequenced from the 5' direction. SSH libraries were cloned non-directionally and had relatively short insert lengths (200–500 base pairs on average); frequently single-pass sequencing was sufficient to generate the entire insert sequence. EST clustering was carried out for individual libraries, individual tissues, for all reads from the 3' direction and for the entire sequence collection (all 3', 5' and SSH reads). Data for the clustering runs on sequences for individual libraries and tissues can be accessed from the CGP database (<http://ri/imb.nrc.ca/codgene/index.php>), with statistics from the clustering runs of 3'-only and the entire collection shown in Table 4.

A total of 154,142 reads from the set of 158,877 sequences were clustered due to additional screening criteria applied by the clustering software (PTA (Paracel 2002)). Clustering of the entire sequence set (All Version 2.0) resulted in 23,838 contigs and 27,976 singletons. The largest number of ESTs in a contig was 154 (all_v2.0.14.C13, best hit interferon stimulated gene 15, *G. morhua*), with an overall average of 5.27 ESTs per contig. The total number of putative transcripts (number of contigs plus

number of singletons) identified in All Version 2.0 was 51,814. The average trimmed length for ESTs in this clustering run was 563 bp. This is slightly lower than that for the clustering of 3' reads alone (591 bp average length) due to the shorter trimmed read lengths for sequences generated from the SSH libraries which were not included in the 3' clustering. For 3' reads only, 97,976 reads were included in EST clustering, resulting in 13,448 contigs, 21,746 singletons and a total of 35,194 for the number of putative transcripts. This lower estimate for the number of transcripts in Atlantic cod is probably a better estimate of the actual number of genes in this species, as the overall assembly contains many non-overlapping sequences for clones and transcripts, including both 5' and 3' reads from sequencing clones, but also potential internal regions of transcripts generated through SSH sequencing.

Annotation of Sequences

The number of putative transcripts which have a significant BLASTX hit (using 1×10^{-5} as the *E* value cut-off) with an existing entry in the NCBI nr database was 15,873 (30.6%) from the complete sequence set (All Version 2.0) and 8,628 (24.5%) from the 3' reads only. Both datasets are biased for

Table 3 SSH libraries created for *Gadus morhua*

Library name	No. of fish treated	No. of fish control	Family
gmnlstfc	18	12	NL YC1 F1 family 32
gmnlstic	18	12	NL YC1 F1 family 32
gmnlkfc	18	12	NL YC1 F1 family 32
gmnlkric	18	12	NL YC1 F1 family 32
gmnlkfta	32	32	NL YC1 F1 family 4
gmnlkrta	32	32	NL YC1 F1 family 4
gmnlfta	32	32	NL YC1 F1 family 4
gmnlrtta	32	32	NL YC1 F1 family 4
gmnlmfta	32	32	NL YC1 F1 family 4
gmnlbfc	24	16	NL YC1 F1 family 32
gmnlbric	24	16	NL YC1 F1 family 32
gmnlkfas	20	20	NL YC1 F1 family 32
gmnlkras	20	20	NL YC1 F1 family 32
gmnlpfas	20	20	NL YC1 F1 family 32
gmnlprsa	20	20	NL YC1 F1 family 32
gmnlpfic	20	20	NL YC1 F1 family 32
gmnlpric	20	20	NL YC1 F1 family 32
gmnlstfas	20	20	NL YC1 F1 family 32
gmnlstas	20	20	NL YC1 F1 family 32

The numbers of fish contributing to each library are shown, together with the program and family details. In the “forward” SSH libraries (i.e. enriched for transcripts that were up-regulated by the treatment), the treated fish samples were used as the tester and the control samples were used as the driver. In the “reverse” SSH libraries (i.e. enriched for transcripts that were down-regulated by the treatment), the control fish samples were used as the tester and the treated samples were used as the driver as previously described (Rise et al. 2008). All fish were juveniles enrolled in the NL CGP breeding program.

the 3' region of transcripts, with the low number of significantly similar sequences likely to be due to the high representation of 3' UTRs within the sequence generated.

The complete sequence set was annotated using an automated pipeline based on AutoFACT (Koski et al. 2005). This performs sequence searches across multiple

databases, combining the results to give a consensus annotation. AutoFACT uses a pipeline for automated annotation that attempts to use corroborating information from multiple sources before assigning an annotation to a sequence. It will also give informative hits greater priority than non-informative hits in the final annotation produced.

Table 4 EST assembly summary statistics

	All Version 2.0	3' reads only
Number of good sequences	154,142	97,976
Average trimmed EST length (bp)	563	591
Number of contigs	23,838	13,448
Number of singletons	27,976	21,746
Number of putative transcripts	51,814	35,194
Maximum no. of ESTs per contig	154	83
Average no. of ESTs per contig	5.27	5.66
Number of putative transcripts with		
Significant BLASTX hits	15,873	8,628
No significant BLAST hits	35,941	26,566
Percentage with no significant BLAST hits	69.37	75.48
Number of contigs containing		
2 ESTs	9,618	4,415
3 ESTs	3,993	2,310
4 ESTs	2,423	1,444
5-10 ESTs	5,194	3,558
11-20 ESTs	1,886	1,309
21-30 ESTs	452	279
31-50 ESTs	221	121
>50 ESTs	51	12

The number of putative transcripts is defined as the number of contigs plus the number of singletons. A score of 1×10^{-5} was used as the *E* value cut-off for BLAST

All singletons and contigs were annotated using this approach. The 3'-only data were used primarily for SNP identification. A bit score of 40 was used as the cut-off value for AutoFACT annotation, allowing the comparison of equivalent hits from different databases. AutoFACT associated informative annotation to 16,258 putative transcripts, with a further 2,090 putative transcripts matching database entries that have no functional annotation attached.

In total, 2,287 putative transcripts can be assigned to KEGG pathways, with the most frequent assigned name being “ribosome” (312 sequences), “oxidative phosphorylation” (145) and “ubiquitin mediated proteolysis” (114). Similar results are obtained after comparison to the COG dataset, with 2,315 sequences assigned to COG categories, the most frequent meaningful categories being “translation, ribosomal structure and biogenesis” (382 sequences), “posttranslational modification, protein turnover, chaperones” (344) and “energy production and conversion” (136). Sequences have also been assigned to Gene Ontology (GO) categories as follows: 1,047 in the cellular component category with the most frequent being “membrane” (161), “intracellular” (161) and “nucleus” (123); 2,801 in the molecular function category, with “protein binding” (139), “catalytic activity” (131) and “zinc ion binding” (120) the most frequent; and 1,366 for the biological process category, with “transport” (92), “immune response” (91) and “proteolysis” (84) being the most frequent. Annotated sequences can be accessed from the project database (<http://ri.imb.nrc.ca/codgene/index.php>).

SNP Detection and Analysis

Contigs generated from clustering of 3' reads only were used for SNP identification to reduce the likelihood of designing amplicons for SNP assays in spliced regions, and to prevent the over-sampling from a single clone that would occur if both 3' and 5' sequences had regions of overlap within a contig. A SNP identification pipeline incorporating PolyPhred (Nickerson et al. 1997) was developed, with custom Perl scripts used to extract information regarding contig coverage, predicted SNPs and the proportion of contributing sequences harbouring each variant (Hubert et al. 2009). SNPs identified in this process were grouped into categories based on the ratio of sequences having each variant, making the assumption that sequences from two independent clones in a contig that originated from the same allele of the same individual would represent a rare event. An ideal situation for a predicted frequent SNP (PF-SNP) was expected to consist of a contig with many contributing sequences over the region analysed where the two different versions of the SNP were present in a 50:50 ratio. Ultimately, the criteria chosen for selection of the

SNPs included a minimum of four read coverage at the position of the SNP within a contig, and a minor allele frequency greater than 25%, which ensured that each version of the SNP was sampled at least twice. This led to the identification of 4753 PF-SNPs. PF-SNPs were chosen for SNP assay development as they were likely to be more frequently variable within family crosses in the breeding program, thus allowing their placement on a genetic map and their potential association with QTL. SNPs were identified as predicted rare SNPs (PR-SNPs) if the minor allele was observed infrequently within a contig, often a single read with a sequence variant, and these have been excluded from SNP assay development at this stage. We also treated PR-SNPs as potential artefacts as there is an increased likelihood that they arose from polymerase errors or sequence mis-calls rather than representing real variation. A small subset of PF-SNPs was confirmed as variable using a low throughput pipeline (Hubert et al. 2009) with a larger set subsequently used for high throughput genotyping using the Illumina GoldenGate platform.

The sequences used for SNP identification originated from two year classes of fish enrolled in the selective breeding programs, NB YC1 and NL YC2. Fish from which tissues were extracted for cDNA library construction included adults which contributed eggs and sperm for the generation of F1 individuals, F1 progeny (embryos, larvae and juveniles) and adults from the same populations as the wild fish used as broodstock, but which did not make a genetic contribution to the breeding programs. The assembly of the 3' reads used for SNP discovery consisted of approximately 58% sequences originating from NB YC1 and 41% from sequences originating from NL YC2 (Table 2), with 1% of sequences from other sources. Because we had identified SNPs based on their frequency in sequences generated from NB YC1 and NL YC2, we expected our SNP set to be most variable in these populations and also to show the highest values for observed heterozygosity and minor allele frequency. Table 5 shows the results of high throughput GoldenGate genotyping for our selected SNPs on several populations of fish, collected from Canada and Europe. The fish genotyped from NB YC1 and NL YC2 were from the same populations as the fish that were used as parents in those year classes, but were not used as parents themselves so did not contribute to the sequence data. Table 6 shows values for observed heterozygosity and minor allele frequency for a small number of SNPs selected from the larger set in fish from Cape Sable and Bay Bulls populations.

The population used to produce NB YC1 showed the greatest number of polymorphic SNPs (71%). For the population used to generate NL YC2, 69% of SNPs were polymorphic. Two other Canadian populations were tested which contributed to two additional year classes enrolled in

the CGP breeding programs, NB YC2 from Georges Bank in the Gulf of Maine, and NL YC3 from Smith Sound, NL. These populations also showed high levels of polymorphism (71% and 70% respectively). Fish isolated in European waters showed a slightly lower level of polymorphism, with fish from Iceland, Ireland and Norway having polymorphism levels of 61%, 53% and 56%, respectively.

Selection of Sequences for Microarray Fabrication

Putative transcripts have been selected from All Version 2.0 for inclusion on a 20,000 element, 50-mer oligonucleotide microarray. Initial criteria for selection attempted to maximise the inclusion of sequences with informative annotation (both contigs and singletons), followed by sequences with similarity to database entries of unknown function. The remaining sequences were selected based on criteria such as known sequence directionality (for design of sense oligonucleotides) and degree of representation, including contigs with a higher number of contributing sequences in preference to contigs with low depth of coverage, or singletons. A few of the deeper contigs containing SSH sequences alone (which were cloned non-directionally), but lacking significant BLAST hits, were included as both sense and anti-sense sequences, in the hope that they represented genes important in response to stimuli such as antigen injection or heat shock. All other contigs and singletons composed entirely of SSH-identified ESTs, but having significant sequence similarity (e.g. *E* values less than 1×10^{-5}), were included in the microarray probe design pipeline as they could be oriented using their BLAST hit. Selection of sequences included on the microarray will be discussed in detail elsewhere (M. Booman, personal communication). However, many of these sequences were also in the set used for SNP detection. A subset of oligonucleotides that were included on the microarray has been designed from sequences for which an informative SNP was also developed. In total, 1,391 oligonucleotides included on the 20,000 element microarray fall into this

category, and some examples taken from the larger set are shown in Table 6. Of these, 605 oligonucleotides have associated AutoFACT annotation, with 567 having informative annotation and 38 hits annotated as “unassigned protein”. The remaining 786 (56%) remain “unclassified” by AutoFACT as the sequences from which they were derived currently have no significant similarities in any of the databases searched. Many of the annotated sequences have similarities to structural or housekeeping proteins, but some have similarity to proteins involved in pathways of particular interest, such as immune response, thermal adaptation or reproduction. Two examples of functional categories identified from the larger SNP set, ribosomal subunits and proteins potentially involved in immunological or stress response pathways, are shown in Table 6.

Discussion

Species-specific genomics projects have frequently been carried out in an iterative fashion, with an initial sequencing phase followed by a marker development phase, followed by the subsequent application of sequences and markers in downstream analyses such as microarray development and QTL identification (Rise et al. 2004; Snelling et al. 2005; Moen et al. 2008). The genomic information for a single species has often been developed by many groups of researchers without overall coordination, leading to the generation of sequences from strains or varieties that are not in current commercial use. Thus sequence information cannot be directly mined for markers relevant in commercial production, resulting in a requirement for additional re-sequencing to develop markers in varieties or strains used in industry. We have designed an integrated approach (Fig. 1) in developing genomics tools for Atlantic cod, such that the information generated is directly and immediately applicable in commercial cod breeding.

Our sequencing approach was designed to maximise gene identification, to ensure that genes of relevance in

Table 5 SNP characteristics

Population	No. of fish genotyped	Polymorphic SNPs	Sequence contribution of population to SNP assembly
NB YC1 (Cape Sable)	23	71%	58%
NB YC2 (Georges Bank)	24	71%	N/A
NL YC2 (Bay Bulls)	23	69%	41%
NL YC3 (Smith Sound)	23	70%	N/A
Iceland	26	61%	N/A
Ireland	15	53%	N/A
Norway	25	56%	N/A

The percentage of SNPs tested that were polymorphic in different populations of fish is shown. Only NB YC1 and NL YC2 fish provided samples from which the sequence collection was generated. N/A = not applicable.

Table 6 Examples of sequences which were used in microarray construction that also harbour a SNP

Contig name	Best Hit in NCBI nr database			<i>E</i> value	SNPs		
	Accession number	Annotation	Species		SNP name	Cape Sable	Bay Bulls
Ribosomal							
all_v2.0.609.C4	gb ACQ58145.1	39S ribosomal protein L30, mitochondrial precursor	<i>Anoplopoma fimbria</i>	1.00E-11	cgpGmo-S198	0.09/0.04	0.30/0.15
all_v2.0.4842.C1	gb ACO09691.1	39S ribosomal protein L32, mitochondrial precursor	<i>Osmerus mordax</i>	7.00E-46	cgpGmo-S967a	0.61/0.43	0.52/0.43
all_v2.0.9851.C1	gb ACI67749.1	39S ribosomal protein L52, mitochondrial precursor	<i>Salmo salar</i>	2.00E-35	cgpGmo-S878	0.43/0.22	0.48/0.37
all_v2.0.560.C2	gb ACH70774.1	ribosomal protein S7	<i>Salmo salar</i>	1.00E-101	cgpGmo-S1917	0.30/0.20	0.26/0.22
all_v2.0.3432.C1	gb ACN10030.1	60S ribosomal protein L10	<i>Salmo salar</i>	1.00E-119	cgpGmo-S105	0.17/0.13	0.30/0.15
all_v2.0.1527.C1	gb ACN10349.1	60S ribosomal protein L18a	<i>Salmo salar</i>	1.00E-90	cgpGmo-S1730	0.35/0.30	0.26/0.22
all_v2.0.4345.C1	gb ACI67287.1	60S ribosomal protein L27	<i>Salmo salar</i>	9.00E-65	cgpGmo-S474	0.22/0.20	0.26/0.26
all_v2.0.14203.C1	gb ACN10033.1	60S ribosomal protein L38	<i>Salmo salar</i>	5.00E-25	cgpGmo-S916	0.35/0.26	0.39/0.24
all_v2.0.983.C1	gb ACO09602.1	60S ribosomal protein L8	<i>Osmerus mordax</i>	1.00E-140	cgpGmo-S357	0.43/0.22	0.39/0.32
all_v2.0.64.C1	dbj BAF98661.1	Ribosomal protein L13a	<i>Solea senegalensis</i>	1.00E-91	cgpGmo-S1519	0.35/0.30	0.35/0.35
all_v2.0.603.C1	dbj BAF45898.1	Ribosomal protein S10	<i>Solea senegalensis</i>	4.00E-69	cgpGmo-S2265	1.00/0.50	1.00/0.50 ^a
all_v2.0.545.C7	gb ACO09841.1	40S ribosomal protein S16	<i>Osmerus mordax</i>	5.00E-67	cgpGmo-S649a	0.30/0.20	0.52/0.26
all_v2.0.3134.C1	ref NP_001134397.1	40S ribosomal protein SA	<i>Salmo salar</i>	1.00E-139	cgpGmo-S699b	0.22/0.11	0.09/0.04
Immune/Stress							
all_v2.0.2956.C2	gb ACO09614.1	Complement component 1 Q subcomponent-binding protein	<i>Osmerus mordax</i>	1.00E-115	cgpGmo-S1200	0.52/0.48	0/0
all_v2.0.4174.C1	gb ACO14444.1	Cystatin-F precursor	<i>Esox lucius</i>	1.00E-27	cgpGmo-S296	0.04/0.06	0.13/0.06
all_v2.0.2115.C1	gb ACN10355.1	CXC chemokine receptor type 4	<i>Salmo salar</i>	1.00E-59	cgpGmo-S525	0.43/0.26	0.35/0.17
all_v2.0.2319.C4	gb AAF72567.1	Immunoglobulin D heavy chain constant region variant a	<i>Gadus morhua</i>	3.00E-33	cgpGmo-S1542	1.00/0.50	1.00/0.50 ^a
all_v2.0.4114.C1	ref XP_001923855.1	Similar to interleukin 12 receptor beta 2.b	<i>Danio rerio</i>	1.00E-23	cgpGmo-S946	0.35/0.48	0.65/0.37
all_v2.0.1550.C1	gb ACO14448.1	Macrophage migration inhibitory factor	<i>Esox lucius</i>	1.00E-41	cgpGmo-S2107	0.52/0.39	0.30/0.24
all_v2.0.8001.C1	ref XP_001894282.1	T-cell receptor beta chain ANA 11	<i>Brugia malayi</i>	2.00E-07	cgpGmo-S1484	0.48/0.28	0.48/0.33
all_v2.0.682.C1	gb ACN10899.1	T-complex protein 1 subunit alpha	<i>Salmo salar</i>	1.00E-106	cgpGmo-S772	0.30/0.28	0.52/0.43
all_v2.0.3543.C1	ref NP_001133482.1	T-complex protein 1 subunit delta	<i>Salmo salar</i>	1.00E-107	cgpGmo-S2247	1.00/0.50	1.00/0.50 ^a

The contig name in the All Version 2.0 assembly is given, together with the accession number and annotation of the best BLAST hit in the NCBI nr database. The SNP name for a SNP identified on that contig is shown, together with values for the observed heterozygosity and minor allele frequency for that SNP in fish collected from Cape Sable, Nova Scotia and Bay Bulls, Newfoundland.

^a Indicates putative SNPs that may represent variation between different genes rather than different alleles of the same gene

aquaculture were present in our collection, and to allow identification of markers present in two CGP breeding programs for use in linkage mapping and QTL identification. Normalized cDNA libraries were constructed from major tissues, including liver, heart, brain, gill, skeletal muscle, digestive tissue (pyloric caecae, anterior stomach, anterior intestine and middle intestine), spleen, head kidney, blood, ovary and testis (Table 1). Four developmental stages were also used in library construction, including embryo, larvae, juvenile (F1 progeny within the breeding programs) and adult (either parents, or fish from the same populations as the parents enrolled in the breeding

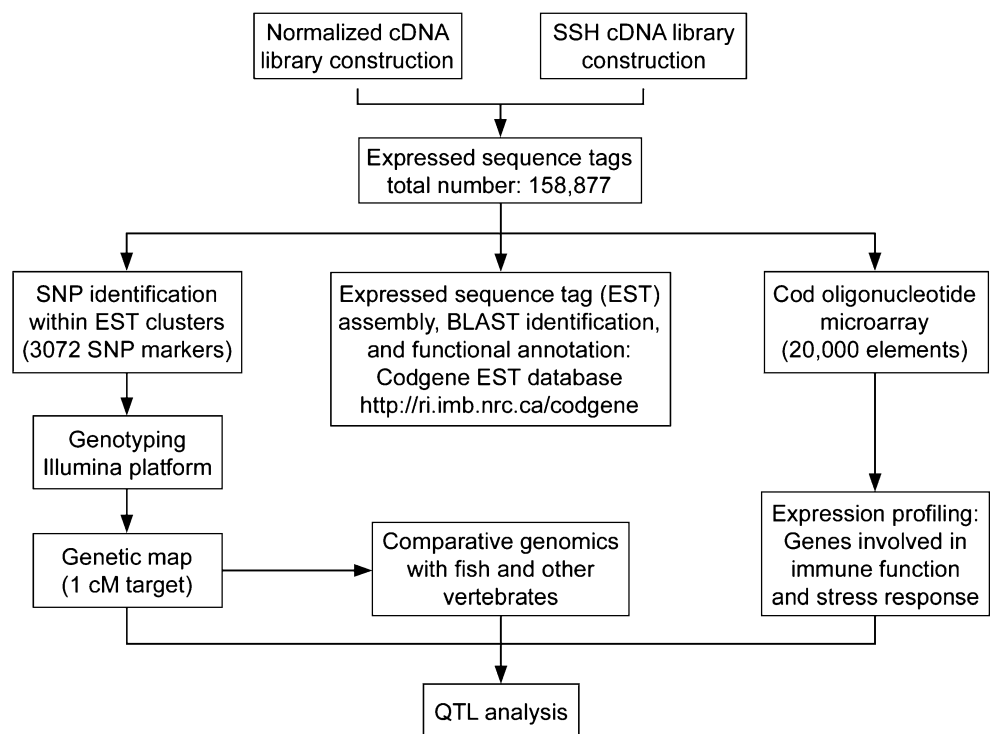
programs). Some of these libraries were selected to target genes involved in specific processes such as reproduction (liver, ovary, testis, embryo) and genes contributing to the immune response (head kidney, spleen, blood). In addition, several libraries were constructed from tissues collected from fish that had been subjected to biotic or abiotic stressors, including exposure to elevated temperature (liver, brain, gill), to the viral mimic pIC (head kidney, spleen, pyloric caecae, blood), to nodavirus (liver) and to the bacterial pathogen atypical *A. salmonicida* (head kidney, blood). Multiple individuals were used per library for the purpose of marker development, to ensure that the sequence

contribution from a single individual was low, and that the variation captured originated from multiple fish.

As the ability to identify highly up-regulated transcripts is lost through use of normalization protocols, we also constructed a series of SSH libraries, with a focus on identifying genes involved in the immune response, or those that respond to temperature stress. Although these libraries were constructed from multiple individuals, each library was constructed from a single family of fish, and therefore any sequence variation captured represented the variation present in only two individuals i.e. the parents giving rise to that cross (Table 3). Although we would expect to find many differences between two unrelated individuals, it was possible that alleles would be identified in this set of sequences that were rare within the population as a whole. Therefore the SSH sequences were not included for marker discovery using the automated pipeline. The SSH library construction protocol uses restriction enzyme digestion with *RsaI* followed by non-directional cloning (Diatchenko et al. 1996), giving rise to sequences of shorter read length where directionality could only be inferred if they assembled with a sequence of known directionality from the collection of normalized reads, or if an informative sequence similarity was found after database searching. However, in general, similarities were identified at higher frequency for SSH libraries when compared to normalized libraries, as the former often represented an internal coding cDNA fragment, whereas the latter contained a high proportion of non-coding 3' sequence.

A set of microsatellite markers was identified in this sequence at an early stage of production (Higgins et al. 2009), but it was necessary to accumulate sufficient depth of coverage within contigs before SNP detection could commence. Identification of putative SNPs was carried out on contigs generated through clustering and assembly of reads generated from the 3' end of normalised libraries. The 5' reads were not used to prevent over-representation of these clones in the sequence used for SNP discovery. SSH sequences were not used because of the limited genetic variation captured in these libraries as described above. The pipeline used for SNP identification has been described in detail elsewhere (Hubert et al. 2009). This study describes the SNPs with respect to the sequences and libraries from which they were generated, and in the context of the CGP breeding programs. Two year classes contributed to sequence generation, from which the SNPs were isolated; NB YC1 and NL YC2. The contribution of each year class to the EST assembly used for SNP discovery was 58% and 41%, respectively. With respect to NB YC1 and NL YC2, the individuals used for genotyping were taken from the same populations as the parents used for generation of those year classes, rather than the parents themselves to avoid over-estimating the levels of polymorphism in those populations. Because the SNPs had been selected based on their predicted high variability in those year classes, we expected to see the greatest levels of both polymorphism and minor allele frequencies in those populations, with NB YC1 predicted to have the higher value because of its larger contribution to the sequence data. This was found to be

Fig. 1 An overview of the genomics workflow within the CGP. The sequence information has been generated from populations of fish enrolled in selective breeding and is being used to develop high-throughput molecular resources for Atlantic cod. These genomics tools (such as genetic markers and an oligonucleotide microarray) will be applied to dissection of QTL within the family-based breeding programs (adapted from Rise et al, 2009)



largely the case, although the population used in the generation of NB YC2 (Georges Bank) showed marginally higher levels of polymorphism than NL YC2. However, the two populations used for production of the NB year classes, Cape Sable and Georges Bank, are from the adjacent NAFO subdivisions 4X and 5Y, respectively, so this is not entirely unexpected. Again, the population used for NL YC3 was not used for sequence generation, and had intermediate values for polymorphism when compared with the other Canadian populations. In comparison, three European populations were tested, showing levels of polymorphism ranging from 53% to 61%. Therefore it will be possible to use the SNPs described here to analyse populations across the North Atlantic, although their utility will be greatest for Canadian populations.

Although it is possible that the differences in the levels of polymorphism observed between Canadian and European fish could be due to a lower level of genetic diversity for this species in European populations, it is unlikely, and the reduction observed is most probably due to ascertainment bias related to the origin of the sequence from which the SNPs were identified. It is therefore likely to result from the limited gene flow between remote populations. The polymorphic loci shared between all populations across the North Atlantic represent a valuable pool of potentially ancient, conserved polymorphisms that could find widespread application within population studies and breeding programs.

As we aimed to capture sequence diversity at the same time as maximising gene identification, parameters used for clustering and assembly allowed sequences with a small number of sequence differences to be incorporated within a contig. In addition, because a large number of individuals were being pooled in the construction of each library, we expected multiple alleles to be present for each gene within the sequence collection. However, this has also made it difficult to determine by bioinformatics means alone whether a contig contains different alleles of the same gene, multiple, closely related genes, or a combination of these two possibilities. We have tested this by genotyping the set of SNPs generated in several CGP families to determine whether SNPs are inherited in a manner consistent with Mendelian segregation. The observation that both parents and progeny from a single cross appear to be heterozygotes is diagnostic for inter-gene differences rather than allelic variants. From these results, we estimate that 4% of contigs represent multiple genes assembled together, rather than the desired result in our assembly such that one contig contains only alleles of a single gene. The ideal substrate for sequencing to facilitate correct gene and allele identification would be a doubled haploid fish (Streisinger et al. 1981). It should be possible to generate doubled haploid individuals in Atlantic cod that are

homozygous at all loci, and this may be an appropriate approach to take in sequencing the genome of this species. Nevertheless, the incidence of multiple genes similar enough to assemble together in a single contig appears to be reasonably low in our assembly, and a more in-depth study of gene families can dissect out the different members and alleles by further sequencing and genotyping (Borza et al. 2009).

The CGP sequences have also been used to design oligonucleotides for spotting on a microarray (Fig. 1) for use in expression profiling. The focus of this effort has been to include sequences having informative annotation, whether they are represented once, or multiple times in our dataset. However, our selection strategy also incorporates sequences that are frequently represented in the CGP sequence collection, and this has ensured that the majority of the sequences used for SNP development are also present on the microarray. Because the sequences originated from tissues taken from fish which had been exposed to a variety of stressors such as immunogenic stimuli (e.g. exposure to bacterial antigens or a viral mimic) or thermal stress, or for several developmental stages, such as eggs, larvae and steps in the reproductive cycle in both sexes, genes involved in these processes are likely to have been expressed, and therefore will be present in our sequence collection. However, because the sequence generated is biased for the 3'UTR to facilitate clustering of individual transcripts and to aid in SNP identification, many of these sequences are currently without informative annotation. Nevertheless, we have been successful in identifying many transcripts where sequence similarity allows us to predict that they function in processes such as reproduction or the immune response. Many of these are represented on the microarray but, in addition, harbour a SNP that has been shown to exhibit polymorphism within the two populations used to generate fish for NB YC1 and NL YC2 (Cape Sable and Bay Bulls, respectively) within the CGP breeding programs. These transcripts will be extremely useful in connecting expression patterns with QTL, with the ultimate aim of identifying the causative genes for traits of interest. For example, a transcript up-regulated on stimulation with viral, bacterial and parasite antigens which is located in a QTL interval associated with disease resistance would be prioritised in further studies.

The integrated genomics tools described here, produced from a single sequence dataset that is tightly linked to fish breeding programs being developed for commercial aquaculture, should enable the rapid dissection of traits and the identification of linked markers which can be applied immediately within these programs to accelerate the development of aquaculture for Atlantic cod.

In conclusion, we have developed a significant sequence resource for Atlantic cod, a species previously under-represented in sequence databases. These sequences have

been generated from fish involved in selective breeding and we have been successful in capturing sequence variation in individuals enrolled in these programs. Future work will focus on exploiting this information to identify markers for commercially relevant traits, and on analysis of data from the additional tools that have been developed from this sequence resource, such as a microarray and a high-density genetic map, to enable future high-throughput studies in Atlantic cod.

Acknowledgements This research was supported in part by Genome Canada, Genome Atlantic and the Atlantic Canada Opportunities Agency through the Atlantic Cod Genomics and Broodstock Development Project. A complete list of supporting partners can be found at www.codgene.ca/partners.php. We would like to thank fish husbandry teams at the Joe Brown Aquatic Research Building at the Ocean Sciences Centre of Memorial University, NL, at the Fisheries and Oceans Canada St. Andrews Biological Station, NB and at the Huntsman Marine Sciences Centre, NB, as well as all individuals who were involved in tissue sampling. We also thank Marie-Catherine Tessier and colleagues at the Genome Quebec and McGill University Innovation Centre for high-throughput genotyping.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Bonaldo MF, Lennon G, Soares MB (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 6:791–806
- Borza T, Stone C, Gamperl AK, Bowman S (2009) Atlantic cod (*Gadus morhua*) hemoglobin genes: multiplicity and polymorphism. *BMC Genet* 10:51
- Bowman S, Higgins B, Stone C, Kozera C, Curtis BA, Tarrant Bussey J, Kimball J, Verheul H, Johnson SC (2007) Generation of genomics resources for Atlantic cod (*Gadus morhua*): progress and plans. *Bull Aquac Assoc Can* 105:24–30
- Brown JA, Minkoff G, Puvanendran V (2003) Larviculture of Atlantic cod (*Gadus morhua*): progress, protocols and problems. *Aquaculture* 227:357–372
- Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci U S A* 93:6025–6030
- Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–194
- Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8:175–185
- Feng CY, Johnson SC, Hori TS, Rise M, Hall JR, Gamperl AK, Hubert S, Kimball J, Bowman S, Rise ML (2009) Identification and analysis of differentially expressed genes in immune tissues of Atlantic cod stimulated with formalin-killed atypical *Aeromonas salmonicida*. *Physiological Genomics* 37:149–163
- Gollock MJ, Currie S, Petersen LH, Gamperl AK (2006) Cardiovascular and haematological responses of Atlantic cod (*Gadus morhua*) to acute temperature increase. *J Exp Biol* 209:2961–2970
- Gorodkin J, Cirera S, Hedegaard J, Gilchrist MJ, Panitz F, Jorgensen C, Scheibye-Knudsen K, Arvin T, Lumholdt S, Sawera M, Green T, Nielsen BJ, Havgaard JH, Rosenkilde C, Wang J, Li H, Li R, Liu B, Hu S, Dong W, Li W, Yu J, Staefeldt HH, Wernersson R, Madsen LB, Thomsen B, Hornshøj H, Bujie Z, Wang X, Bolund L, Brunak S, Yang H, Bendixen C, Fredholm M (2007) Porcine transcriptome analysis based on 97 non-normalized cDNA libraries and assembly of 1, 021, 891 expressed sequence tags. *Genome Biol* 8:R45
- Hansen T, Karlsen O, Taranger GL, Hemre G, Holm JC, Kjesbu OS (2001) Growth, gonadal development and spawning time of Atlantic cod (*Gadus morhua*) reared under different photoperiods. *Aquaculture* 203:51–67
- Hayes BJ, Nilsen K, Berg PR, Grindflek E, Lien S (2007) SNP detection exploiting multiple sources of redundancy in large EST collections improves validation rates. *Bioinformatics* 23:1692–1693
- He C, Chen L, Simmons M, Li P, Kim S, Liu ZJ (2003) Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. *Anim Genet* 34:445–448
- Higgins B, Hubert S, Simpson G, Stone C, Bowman S (2009) Characterization of 155 EST-derived microsatellites and validation for linkage mapping. *Mol Ecol Resour* 9:733–737
- Hori TS, Kimball J, Johnson SC, Afonso LOB, Bowman S, Hubert S, Gamperl AK, Rise ML (2010) Heat-shock responsive genes identified and validated in Atlantic cod (*Gadus morhua*) using genomic techniques. *BMC Genomics* 11:72
- Hubert S, Tarrant Bussey J, Higgins B, Curtis BA, Bowman S (2009) Development of single nucleotide polymorphism markers for Atlantic cod (*Gadus morhua*) using expressed sequences. *Aquaculture* 296:7–14
- Hutchinson WF, Carvalho GR, Rogers SI (2001) Marked genetic structuring in localised spawning populations of cod (*Gadus morhua*) in the North Sea and adjoining waters, as revealed by microsatellites. *Mar Ecol Prog Ser* 223:251–260
- Kan Z, Gish W, Rouchka E, Glasscock J, States DJ (2000) UTR reconstruction and analysis using genomically aligned EST sequences. *Proc Int Conf Intell Syst Mol Biol* 8:218–227
- Khan RA (2005) Prevalence and influence of *Loma branchialis* (Microspora) on growth and mortality in Atlantic cod (*Gadus morhua*) in coastal Newfoundland. *J Parasitol* 91:1230–1232
- Koop BF, von Schalburg KR, Leong J, Walker N, Lieph R, Cooper GA, Robb A, Beetz-Sargent M, Holt RA, Moore R, Brahmabhatt S, Rosner J, Rexroad CE 3rd, McGowan CR, Davidson WS (2008) A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics* 9:545
- Koski LB, Gray MW, Lang BF, Burger G (2005) AutoFACT: an automatic functional annotation and classification tool. *BMC Bioinformatics* 6:151
- Labate JA, Baldo AM (2005) Tomato SNP discovery by EST mining and resequencing. *Mol Breeding* 16:343–349
- Lijavetzky D, Cabezas JA, Ibanez A, Rodriguez V, Martinez-Zapater JM (2007) High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* 8:424
- Moen T, Hayes B, Baranski M, Berg PR, Kjøglum S, Koop BF, Davidson WS, Omholt SW, Lien S (2008) A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. *BMC Genomics* 9:223
- Myers RA, Hutchings JA, Barrowman NJ (1997) Why do fish stocks collapse? The example of cod in Atlantic Canada. *Ecol Appl* 7:91–106
- Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions

- using fluorescence-based resequencing. *Nucleic Acids Res* 25:2745–2751
- PARACEL (2002) PTA: Paracel transcript assembler user manual. Paracel, Inc, Pasadena, CA
- Park KC, Osborne JA, Tsoi SC, Brown LL, Johnson SC (2005) Expressed sequence tags analysis of Atlantic halibut (*Hippoglossus hippoglossus*) liver, kidney and spleen tissues following vaccination against *Vibrio anguillarum* and *Aeromonas salmonicida*. *Fish Shellfish Immunol* 18:393–415
- Patel S, Korsnes K, Bergh O, Vik-Mo F, Pedersen J, Nerland AH (2007) Nodavirus in farmed Atlantic cod *Gadus morhua* in Norway. *Dis Aquat Organ* 77:169–173
- Ramsey JS, Wilson AC, De Vos M, Sun Q, Tamborindeguy C, Winfield A, Malloch G, Smith DM, Fenton B, Gray SM, Jander G (2007) Genomic resources for *Myzus persicae*: EST sequencing, SNP identification, and microarray design. *BMC Genomics* 8:423
- Rise ML, Hall J, Rise M, Hori T, Gamperl AK, Kimball J, Hubert S, Bowman S, Johnson SC (2008) Functional genomic analysis of the response of Atlantic cod (*Gadus morhua*) spleen to the viral mimic polyriboinosinic polyribocytidylic acid (pIC). *Dev Comp Immunol* 32:916–931
- Rise ML, von Schalburg KR, Brown GD, Mawer MA, Devlin RH, Kuipers N, Busby M, Beetz-Sargent M, Alberto R, Gibbs AR, Hunt P, Shukin R, Zeznik JA, Nelson C, Jones SR, Smailus DE, Jones SJ, Schein JE, Marra MA, Butterfield YS, Stott JM, Ng SH, Davidson WS, Koop BF (2004) Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Res* 14:478–490
- Rise ML, Gamperl AK, Johnson SC, Robinson A, Culver K, Booman M, Borza T, Garber A, Hubert S, Tosh J, Afonso L, Boyce D, Hamoutene D, Lush L, Neil S, Symonds J, Butts I, Feng CY, Hori T, Pérez-Casanova J, Shine E, Hall JR, Higgins B, Kimball J, Simpson G, Stone C, Tarrant Bussey J, Elliot J, Moir J, Nardi G, Powell F, Walsh A, Rise M, Trippel EA, Bowman S (2009) The Atlantic Cod Genomics and Broodstock Development Project. *Bull Aquac Assoc Can* 107-3:21-30
- Rodger HD, Colquhoun DJ (2008) Clinical vibriosis in farmed Atlantic cod (*Gadus morhua*) in Ireland. *Vet Rec* 162:94–95
- Rodriguez-Tovar LE, Wadowska DW, Wright GM, Groman DB, Speare DJ, Whelan DS (2003) Ultrastructural evidence of autoinfection in the gills of Atlantic cod *Gadus morhua* infected with *Loma* sp. (phylum Microsporidia). *Dis Aquat Organ* 57:227–230
- Rose GA (2007) Cod: The ecological history of the North Atlantic fisheries. Breakwater Books, St. John's, Newfoundland
- Rosenlund G, Skretting M (2006) Worldwide status and perspective on gadoid aquaculture. *ICES J Mar Sci* 63:194–197
- Ruzzante DE, Taggart CT, Cook D, Goddard S (1996) Genetic differentiation between inshore and offshore Atlantic cod (*Gadus morhua*) off Newfoundland: Microsatellite DNA variation and antifreeze level. *Can J Fish Aquat Sci* 53:634–645
- Snelling WM, Casas E, Stone RT, Keele JW, Harhay GP, Bennett GL, Smith TP (2005) Linkage mapping bovine EST-based SNP. *BMC Genomics* 6:74
- Streisinger G, Walker C, Dower N, Knauber D, Singer F (1981) Production of clones of homozygous diploid zebra fish (*Brachydanio rerio*). *Nature* 291:293–296
- Symonds J, Garber A, Puvanendran V, Robinson A, Neil S, Trippel E, Walker S, Boyce D, Gamperl K, Lush L, Nardi G, Powell F, Walsh A, Bowman S (2007) Family-based Atlantic cod (*Gadus morhua*) broodstock development. *Bull Aquac Assoc Can* 105:39–46
- Zhulidov PA, Bogdanova EA, Schcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA, Shagin DA (2004) Simple cDNA normalization using kamchatcka crab duplex-specific nuclease. *Nucleic Acids Res* 32:e37