R E P O R T

Evaluation of the idea to use cod as a species for genome sequencing and functional genomic studies in Norway

February 2004

Contents

Foreword

Genome projects are resource intensive tasks with potentially positive impact both on the competence of the research environments performing the science and on the advancement of industries based on the species which genomes are being studied. The salmon genome has been studied in several internationally linked projects over many years, and since 2000 in the Norwegian "*Salmon Genome Project*" funded by The Research Council of Norway. In 2002, a collective initiative was taken by the research environments in Bergen and Tromsø to promote a project on genome sequencing and functional genomic studies in cod.

To evaluate this idea in the context with ongoing activities and strategic interests for Norway, The Research Council and The ministry of Fisheries assembled a group of experts which was given mandate to perform this task during the autumn 2003. The results of the group's work are presented in this report.

The Research Council and The Ministry of Fisheries thanks the group for its work.

Karin Refsnes
Executive Director, Dr.philos
Division for Strategic Priorities
The Research Council of Norway.

Oslo, February 2004.

Group members and mandate

In a letter of 11th July 2003 an expert group was appointed the task to evaluate the idea to use cod as a species for genome sequencing and functional genomic studies. The group was provided with background material as shown in appendix 1.

The members were:
Professor Leif Andersson, Dept. of Medical Biochemistry and Microbiology, Uppsala University, Sweden (leader of the group).
Chief Scientist Christian Bendixen, Afd. for Husdyravl og Genetik, Ministeriet for Fødevarer, Landbrug og Fiskeri, Tejle, Denmark.
Professor Daniel Chourrout, Sars International Centre for Marine Molecular Biology, Bergen, Norway.
Dr. Anna Kristin Danielsdottir, The Population Genetic Laboratory, Marine Research Institute Reykjavik, Iceland.
Dr.scient Guri Eggset, NorInnova AS, Tromsø, Norway.
Nesteleder Knut Hjelt, Fiskeri og Havbruksnæringens Landsorganisasjon, Trondheim, Norway.
Professor Hans Prydz, Bioteknologisenteret, Universitetet i Oslo, Norway.

Adviser Dr. Steinar Bergseth, The Research Council was secretary for the group.
Adviser Dr. Sigve Nordrum, The Ministry of Fisheries was participating as an observer.

The group had the following mandate:
As a background for the evaluation these general guidelines should be followed:

- A genome research activity shall build competence about the chosen species' biology by generating knowledge about its genome (functional genomics / post genomics).
- The marine research environments in Bergen and Tromsø shall be strengthened by utilisation of the national and international competence in this area.

The evaluations shall be coordinated with the ongoing activities on salmon and other species and be seen in relation to the eventual funding of further activities on these. Benefits of a marine genomic project in relation to the priorities from The Ministry of Fisheries on cod aquaculture shall also be part of the evaluation.

The report is based on the competence in the group, the supplied documents and discussions with the stakeholders and within the group.

Christian Bendixen Daniel Chourrout Anna Kristin Danielsdottir

Guri Eggset Knut Hjelt Hans Prydz

Leif Andersson (leader) Steinar Bergseth (secretary)

Uppsala/Tjele/Bergen/Reykjavik/Tromsø/Trondheim/Oslo
February 2004

Organisation of the work

The group had its first meeting 16.09.2003 and met thereafter two times to have discussions with the relevant research environments in Bergen and Tromsø and with the same in Oslo and Ås. The meeting agendas are appendixes to the report.

Report recommendations and conclusions are supported by all members of the group.

Summary

The evaluations committee strongly recommends that marine molecular biology should be strengthened in Norway. This is well justified since (i) the marine environment is of outmost importance for Norway, (ii) Norway has a large oil industry that has some negative effects on the marine environment, and (iii) Norway has a strong aquaculture industry.

However, the evaluation group has come to the conclusion that neither the cod nor the salmon projects have reached the stage where a full genome sequencing program could be justified. We therefore recommend that the RCN announce a call for proposal for 2-3 years projects or organise a program in marine genome research. Knowledge generation to build competence for utilisation of sequencing information should have high priority. A strong leadership and strong research programs which can lead to full genome sequencing programs and the associated activities at a later stage should be established.

Many of the challenges the cod industry face today could be addressed through functional genomics. Our recommendation is therefore to initiate a strong research program on cod biology and breeding which will constitute the basis for a genome sequencing program at a later stage.

Norway should as soon as possible take initiative to form international consortia of stakeholders that can make a common effort towards genome sequencing of selected marine species (e.g. salmon and cod) within 5 years.

Introduction and background

Basic concepts in genome programs

The genome sequences of many bacteria and eukaryotes have by now been determined completely. There are two major reasons for sequencing the genome of a species. One is that it may advance our general knowledge in basic biology and the other is that it may lead to important practical applications. The major motivation for sequencing a model organism like the zebrafish is that the access to a complete genome sequence facilitates basic research on this species, and the possibility to do comparative genetic studies with other species. In other species, the genome sequence has also been determined to facilitate both the identification of mutations causing different disorders and the development of new diagnostic tools and treatments for diseases.

The evaluation group strongly supports the view that both salmon and cod are important marine species that eventually should be sequenced. The established salmon industry and the emerging cod industry in Norway could benefit significantly from a strong marine genomics program. The access to the complete genome sequence for cod and/or salmon would greatly facilitate all types of genetic studies and can lead to practical applications in fish breeding. It would also facilitate research on disease resistance and the development of new vaccines and other treatments of disease. Thus, the sequencing of the cod and salmon genomes is primarily motivated by the potential practical implications.

However, the determination of the complete genome sequence of a vertebrate species is a major undertaking. A large research community is required to fully utilize the huge amount of information that is generated when sequencing a genome. Prerequisites for a genome sequencing project is a strong leadership, thorough planning on the organisational structure with partnerships, scientific strategies including IPR, budgets, excellent organization and well developed bioinformatics. For instance, the sequencing of the human genome was preceded by about 20 years of research with the aim to establish a genetic and physical map of the human genome. Moreover, thousands of researchers around the world were eager to take advantage of the genome sequence in their research programs. Thus, the huge investment required to sequence the human genome was justified by the major advance it made for human molecular medicine.

This field is still in the early stages of development, and only limited experience in the utilisation of genomic information is available. In the near future (5 – 10 years) this will certainly change and all genomic efforts started up now will therefore be the basis for the phase of the genomics era when genome information can be utilised practically.

A genome project – what is it?
A short version…

Generating the nucleotide sequence of the complete genome involves relatively few steps: DNA isolation, library construction, followed by (high-throughput) shotgun sequencing. This part of the genome research is predominantly carried out by few large sequencing centres well capable of generating 10-100s of millions of basepair sequence per day. However, the raw sequence data must be transformed to biologically relevant information. This process is a challenging task on its own which includes the steps of clustering and assembly of the sequences as well as by a finishing step, closure of the gaps in the sequence, and finally by annotation of genes and other sequence elements.

The general strategy of the genome sequencing itself is characterized by three procedures: The large chromosomal DNA is fragmented into smaller pieces, which are then randomly

sequenced. Finally, after generating redundant sequence data, all reads of the individual fragments are reassembled into a single, overlapping DNA sequence using computer algorithms. Currently, several different strategies exist to sequence large mammalian genomes: "clone-by-clone" shotgun sequencing and the "whole-genome shotgun" sequencing. Both of these approaches were applied in the sequencing of the human genome: The Human Genome Project used clone-by-clone sequencing where bacterial artificial chromosome (BAC) and cosmid clones were sequenced and assembled individually before a global assembly was performed. Celera Genomics utilized the whole-genome shotgun sequencing approach where a large number of random fragments are sequenced and subsequently all sequences are clustered and assembled using no other knowledge than the overlap of the individual sequences. These methods are not exclusive and can be applied together, and such a mixed "hybrid sequencing strategy" was used in the process of sequencing the rat genome combining the advantages of both methods.

For the clustering and assembly step in the genome analysis there is a need of detailed genetic linkage maps and physical maps in the shape of RH-maps (Radiation-Hybrid maps) or as a BAC (Bacterial Artificial Chromosome) scaffolds assembled by fingerprinting. But, even with access to the raw assembled sequence of a genome it is not possible to accurately define the gene complement of the species, since existing gene finding algorithms are prone to error. This means that annotation of the assembled sequence, particularly identifying the structural elements of genes, introns, exons, and regulatory regions, is at present not precise or efficient without large collections of ESTs (Expressed Sequence Tags) obtained by end-sequencing of cDNA clones. EST sequences are thus a valuable intermediate resource in the validation and annotation process of the genome sequence. Importantly, ESTs provide access to gene-specific mapping information which reveals the syntenic relationship between chromosomes among species allowing for comparative gene maps to be established. Furthermore, mapping of ESTs onto BAC clones helps to relate the linkage and RH map to the physical BAC map. Establishing an EST resource also provides a valuable tool for conducting “functional genomics” by cDNA microarray-studies of gene expression.

Because of the complexity of the assembly and annotation process in genome analysis, it is standard that the analysis of a genome is initiated by the establishment of genetic and physical maps. In most cases the genetic linkage maps are generated by the use of highly polymorphic microsatellite markers supplemented with SNP-based (Single Nucleotide Polymorphism) marker information when available. Physical maps can be developed using radiation hybrid mapping and/or in situ hybridisation. Scaffolds of BAC clones (Bacterial Artificial Chromosome) have been assembled by identifying regions of clone overlap, using a process called fingerprinting, thereby creating large contiguous "tiling paths" which are colinear with the chromosomes and instrumental in the final assembly of a coherent genomic sequence.

Cost estimates for genome projects

Estimating the costs for a genome project is a very difficult task, which will be highly dependent upon a number of factors, the most important being the scale (genome size), precision (sequence quality), coverage (how complete the genome sequence is) and available genome information (genetic markers, linkage maps, etc.) when the sequencing starts. Further more a large genome containing a high frequency of repetitive and duplicated sequences are much more difficult to sequence than a small genome with few repetitive and duplicated sequences. One might suggest a percentage of total project cost to each major part of the project (Genomic sequencing 55%, EST sequencing 12%, genetic and physical map 18%, bioinformatics and assembly 15%). These numbers are also up for discussion and will vary with the species as salmon will require a larger percentage of the budget for genomic sequencing and less for ESTs and mapping.

For Norway to build competence and establish an international, strong position in this field a marine genomic funding increasing from a start of 20 mNOK to 100 mNOK annually within 4 years would give payoff both for scientific competence buildup and practical applications. Costs for participation in genome sequencing programs must be calculated separately. A duration of 5 + 5 years is a prerequisite to reach the goals. Table 1 shows some costs for obtaining different genome sequences.

Table 1
Estimated genome size and costs of genome mapping programmes for some actual species.

Species	**Genome size (bill bases)**	**Cost [1)] (million NOK)**
Honeybee	0.270	49
Chicken	1.2	210
Chimpanzee	3.0	210-350
Kangaroo	3.3	700-1050
Dog	2.8	210-350
Silkworm	0.500	105
Cow	3.2	350-400
Cod	1.0	100
Salmon	3.0	300

1) Cod and salmon are estimated by the evaluation group, the others by New Scientist, 2002.

Status on ongoing genomics activities in Norway (a summary)

Cod and the CodGen proposal

The overall impression from our study visit to Bergen and Tromsø was that the ongoing research activities in the field of molecular biology and molecular genetic studies in the cod are very limited. There are several interesting ongoing research projects, for instance on genes of immunological importance, but the total research community in Norway and in the world working with the cod at the molecular level is too small to justify a complete genome sequencing of the cod at present. If the total genome sequence of the cod became available within say 1 year, there would be a very limited number of researchers that could take advantage of this information.

The panel strongly agrees that a cod genome program is well justified but were not convinced that the full genome sequence is required or justified at present. In the CodGen proposal there are three major areas of applications that are used as arguments for the need of a genome sequencing program but it is the view of the panel that a full genome sequence is not required for these applications.

a. Management of wild cod stocks. The applicants would like to develop more genetic markers to be used in genetic studies of wild cod populations. It is a trivial task to develop hundreds of genetic markers with a very limited research budget. So, this can easily be achieved without sequencing the full genome.

b. Monitoring impacts of environmental factors and climatic changes. Here the applicants propose that it should be possible to use gene expression analysis to monitor to which extent cod has been exposed to environmental factors. This would be possible to achieve with a

medium size EST (Expressed Sequence Tag) sequencing project and the subsequent development of cod-specific DNA arrays.

c. Removing biological bottlenecks in cod aquaculture. No clear strategy has been presented how the applicants will use the genome sequence information to remove bottlenecks in cod aquaculture. One of the possible applications that are mentioned is that genome information can be used to identify Quantitative Trait Loci (QTL). This does not initially require a full genome sequencing program. A few hundred genetic markers (~500) will be sufficient to map QTL for various important traits in the cod if suitable pedigree materials are made available. Such QTLs could then be applied in cod breeding programs by marker assisted selection (MAS). Furthermore, a list of commercially important QTLs localized to specific chromosome regions would provide an argument for a full-scale genome sequencing at a later stage, since this would facilitate the identification of the underlying gene.

Thus, activities within the area of cod EST-sequencing, generation of physical and genetic maps for the cod genome as well as QTL-studies should be supported. This would create the foundation for a subsequent more ambitious project with the aim to fully characterize the cod genome. It should be emphasized that the proposed activities to develop genetic markers, genetic maps and EST resources is not a waste of money whether or not the cod genome sequence becomes available a few years from now. These types of resources are, as described above, very important for a genome sequencing program, but also provides access to important tools within the area of functional genomics.

Salmon and the Salmon Genome Project

The salmon genome research has been ongoing for about 10 years. The Salmon Genome Project was started in 2000 with a yearly budget of 10 mNOK. However, a considerable portion of these resources have been used to build up bioinformatics resources in general in Norway. Many of the resources planned to be developed in the Cod Genome Project have already been established in the salmon, such as a large number of genetic markers, a medium density genetic map, a BAC library and EST resources. Thus, good progress has been made to establish resources for genome research in salmon but the genome project does not yet appear to be well integrated with biological research and breeding programs in salmon. An important application of the salmon genome project would be to use genomic tools to find genes associated with important traits in the salmon, such as disease resistance, meat colour, growth and sexual maturity. This knowledge could then be exploited to improve the traits by breeding or other means when the biology is better understood. Some activities in this area were presented to the evaluation group but the research programs had not yet reached the level where many chromosomal regions harbouring genes with interesting phenotypic effects have been identified. This is in sharp contrast to the situation in for instance humans, mice, chicken or zebrafish when it was decided to sequence these genomes. Thus, the conclusion is therefore that there is no immediate need that justifies the huge investment that is required to fully sequence the salmon genome at present. However, continued and increased support for the present salmon genome project and a better integration of the genome project with salmon biology/salmon breeding is strongly recommended.

Other

In addition to the activities mentioned above, Norway has groups with strong experience in performing genomic research on other organisms as listed below:

- Contributions in the Human Genome Sequencing project and steering bodies.

- Sequencing of bacterial genomes.
- Sequencing and genomics on model organisms like the zebrafish and the urochordate Oikopleura
- Bovine genomics including QTL mapping
- Plant genomics and utilisation of the Arabidopsis genome information.

Marine genomics in an industrial perspective

The established salmon industry with a yearly production of 500.000 tons, and the emerging cod industry in Norway, could on the long run benefit significantly from a strong marine genomic program. The present yearly production of cod is about 1000 tons, and a sharp increase is expected the coming years.

With further growth of salmon production in Norway, the industry will probably face significant challenges including interaction water/water quality/organism, conflicts between freshwater performance and salt water performance, so called production-related malfunctions and susceptibility to viral and bacterial infections. Another future challenge concerns the interaction between wild and farmed salmonids, including gene flow and sea lice exposure. There will also be huge challenges in coping with possible infectious diseases, which mainly would be addressed through genetically improved breeding programs and development of new vaccines (based on ecological and microbiological knowledge of infectious agent/host interaction).

Cod is considered to be the most promising volume species in Norwegian aquaculture. At present the lack of basic biological knowledge is hampering the development of the cod aquaculture industry. As a consequence a number of biological bottlenecks still exist throughout the entire chain of cod production in aquaculture. The problems include a high level of spine malfunctions, a high incidence of cataracts in production populations, early sexual maturation, poor flesh quality, and possible interaction with wild populations. Furthermore, development of effective vaccines and other means to improve health conditions will have to be faced in the efforts to make cod an important species in Norwegian aquaculture. In order to meet these problems competence in genetics is of major importance and considerable economic investments will be required.

When selecting breeding populations for cod there is an urgent need for basic knowledge on the genetic variability of cod that will allow optimal selection of breeding populations that has the potential to perform in accordance with the breeding goals.

Many of the challenges the cod industry face today could be addressed through functional genomics. Our recommendation is therefore to initiate a strong research program on cod biology and breeding which will constitute the basis for a genome sequencing program at a later stage.

Possibilities for international cooperation

The interest in cod genome research is mainly found in Norway, Iceland and Scotland, due to their involvements both in fisheries and aquaculture of cod. However, cod aquaculture is also planned in Denmark, Sweden, USA, Ireland, Russia and Spain, and probably the interest in cod genome will increase in the future. Some research is already found in Sweden. Canada, having started haddock farming, may also be interested in cod genome projects. Cooperation within genome-related research projects already exists between Norway – Canada, Norway – Iceland, Norway – Sweden and Norway – Scotland, and would be easy to extend to new projects and partners.

In salmon genome research, Norwegian research groups (the Salmon Genome Project) already cooperate with strong Canadian groups. There is also some on-going genome research on salmonids (mostly trout) in Scotland, and cooperation with Norwegian groups exists.

In the near future, it seems that most of the work within marine genomics has to be financed by national resources as the theme is not prioritized within the 6th EU Framework Programme. Nevertheless, the group is of the opinion that sequencing projects of the complexity encountered when marine fish genomes are aimed at, should primarily be done by international consortia. Due to the fact that Norway has "responsibility" for a huge part of the north Atlantic salmon population and also the wild population of north east Atlantic cod, Norway should take a leading part in such a consortia. The challenges in patenting and users' right (IPR) for the knowledge gained through the scientific findings in such a work must be addressed adequately.

Conclusions

The evaluation group has come to the conclusion that neither the cod nor the salmon projects have reached the stage where a full genome sequencing program could be justified. The strong impression of the evaluation committee was that ongoing research in the field of marine molecular biology in Norway is limited. This implies that there are only few research groups that could really take advantage of a full genome sequence of the cod or the salmon and make interesting biological research of such a resource and/or use the resource for useful practical applications. Therefore, it appears much more appropriate to use the resources required for a full genome sequencing (hundreds of million NOK) to strengthen marine molecular biology in Norway. This process should be strengthened by international recruitments. A strong leadership and know how also needs to be established which can lead full genome sequencing programs and the associated activities at a later stage.

Recommendations

Recommended actions for the near future

The Norwegian government has allocated 7.5 mNOK for cod genome research in 2004. We recommend that the RCN announce a call for proposal for 2-3 years projects or organise a single program in marine genome research. Knowledge generation in the following areas should have high priority:

- Development of large number (>100) of genetic markers (microsatellites or SNPs)
- Construction of a linkage map and its application to identify Quantitative Trait Loci (QTLs) for important traits
- An EST sequencing program and development of cDNA arrays for expression analysis.
- The development of a high-quality cod BAC library
- Any program that use genomic tools to study cod biology, for instance the use of a large number of genetic markers to advance our knowledge concerning genetic diversity between cod populations and their identification.

The projects/program may involve co-operational institutions and scientists, national and international participants experienced within marine genomics, together with coastal research

institutions that aim to build competence within marine genomics to promote strong research programs on cod biology. With the political importance of marine genomics, it is a strong need for a common structure which implements a firm leadership within this field. It should therefore be considered how this best can be achieved. The group envision at least three possibilities:

1. Implement a national steering body with mandate to recommend allocation of responsibilities, cooperative projects and funding within marine genomics.
2. Establish Centres of Excellence within this field (see next paragraph).

Recommendations to strengthen marine molecular biology in Norway

The evaluations committee strongly recommends that marine molecular biology should be strengthened in Norway. This is well justified since (i) the marine environment is of outmost importance for Norway, (ii) Norway has a large oil industry that has some negative effects on the marine environment, and (iii) Norway has a strong aquaculture industry.

We recommend that Centres of Excellence in Marine Molecular Biology will be established after call for proposals that are reviewed by international experts. The centres should have a strong leadership and the funding could be for 5 + 5 years with evaluation after 5 years. The support for each centre could be in the range 10 - 20 mNOK/year. Cod and salmon genomics could be declared as areas of particular importance in the program. We think this will be a constructive way to strengthen marine molecular biology in general. This will allow the different universities and research institutes to organize highly competitive applications that may involve the potential recruitments of senior scientists with an international reputation.

To achieve full genome sequences, Norway should as soon as possible take initiative to form international consortia of stakeholders that can make a common effort towards genome sequencing of selected marine species (e.g. salmon and cod) within 5 years. Central research environments should get the responsibility to work out the consortia in close collaboration with the Research Council and Ministry of Fisheries to ensure anchoring with the financing institutions at an early stage.