Salmon Genome Project (SGP)
Resources

Bjørn Høyheim
Norwegian School of Veterinary Science
Department of Basic Sciences and Aquatic Medicine
Salmon Genome Project (SGP)

- Genetic map
- BAC library/physical map
- EST-sequencing
- Expression profiling
- QTL studies
- Genomic sequencing
- Bioinformatics

www.salmongenome.no
SGP resources
Summary

- **Genetic map for Atlantic salmon**
  - 450 markers on the consensus map, Average 1 marker pr 2-3 cM
  - A bank of approx. 1200 microsatellites

- **BAC library**
  - A well characterized library
  - 18 fold genome coverage
  - Screened for over 200 genes/markers resulting in more than 4000 BACs identified

- **cDNA libraries**
  - 23 libraries consisting of 160 000 gridded clones
  - 68 000 sequences from these
    - Trimmed and clustered
    - 30% annotated

www.salmongenome.no
SGP resources

Summary

- **A bioinformatics infrastructure**
  - Database for sequences
  - Database for genetic maps
  - Pipelines for sequence manipulations
  - Pipelines for clustering/annotation
  - BLAST search locally

- **Microarray chip being developed**
  - 17,500 genes in duplicate (2005)

www.salmongenome.no
Salmon Genome Project (SGP)

- Genetic map
- BAC library/physical map
- EST-sequencing
- Expression profiling
- QTL studies
- Genomic sequencing
- Bioinformatics

www.salmongenome.no
Atlantic salmon
Genetic maps

- www.salmongenome.no
- Markers with descriptions
- Graphical view
- Genotype database
Atlantic salmon
Genetic maps

- www.salmongenome.no
- Markers with descriptions
- Graphical view
- Genotype database
Atlantic salmon
Genetic markers
Atlantic salmon
Genetic markers

<table>
<thead>
<tr>
<th>Restriction enzyme</th>
<th>Insert-size</th>
<th>No. of libraries</th>
<th>Vector</th>
<th>Laboratory</th>
<th>Microsat-name</th>
</tr>
</thead>
<tbody>
<tr>
<td>HaeIII/RsaI/EcoRI/DraI</td>
<td>300-800bp</td>
<td>2</td>
<td>pBlueScript II</td>
<td>NSVS</td>
<td>BHMS</td>
</tr>
<tr>
<td>AluI</td>
<td>300-800bp</td>
<td>2</td>
<td>pBlueScript II</td>
<td>NSVS</td>
<td>Alu</td>
</tr>
<tr>
<td>HaeIII</td>
<td>300-800bp</td>
<td>1</td>
<td>pBlueScript II</td>
<td>NSVS</td>
<td>Hae</td>
</tr>
<tr>
<td>RsaI</td>
<td>300-800bp</td>
<td>2</td>
<td>pBlueScript II</td>
<td>NSVS</td>
<td>Rsa</td>
</tr>
<tr>
<td>AluI/HaeIII</td>
<td>200-600bp</td>
<td>1</td>
<td>pUC19</td>
<td>NUIG</td>
<td>NUIG</td>
</tr>
<tr>
<td>AluI</td>
<td>200-600bp</td>
<td>2</td>
<td>pUC19</td>
<td>NUIG</td>
<td>NUIG</td>
</tr>
<tr>
<td>HaeIII</td>
<td>200-600bp</td>
<td>1</td>
<td>pUC19</td>
<td>NUIG</td>
<td>NUIG</td>
</tr>
<tr>
<td>MboI/HaeIII/AluI/RsaI</td>
<td>200-1200bp</td>
<td>9*</td>
<td>pBlueScript II</td>
<td>UOS</td>
<td>UOS</td>
</tr>
</tbody>
</table>

* These 9 libraries were constructed by separate digesting with the restriction enzymes. Then all the libraries were pooled for the screening and analysis.

Libraries developed during SALMAP
Libraries developed during SGP
Atlantic salmon
Genetic markers
Atlantic salmon
Genetic markers

<table>
<thead>
<tr>
<th>#No</th>
<th>Marker name</th>
<th>GenBank accession</th>
<th>Marker type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AAT-4*</td>
<td>Not known</td>
<td>other</td>
</tr>
<tr>
<td>2</td>
<td>Alu002A</td>
<td>AF271425</td>
<td>microsatellite</td>
</tr>
<tr>
<td>3</td>
<td>Alu002B</td>
<td>AF271425</td>
<td>microsatellite</td>
</tr>
<tr>
<td>4</td>
<td>Alu005</td>
<td>AF271427</td>
<td>microsatellite</td>
</tr>
<tr>
<td>5</td>
<td>Alu008</td>
<td>AF271428</td>
<td>microsatellite</td>
</tr>
<tr>
<td>6</td>
<td>Alu015</td>
<td>AF271431</td>
<td>microsatellite</td>
</tr>
<tr>
<td>7</td>
<td>Alu017</td>
<td>AF271433</td>
<td>microsatellite</td>
</tr>
<tr>
<td>8</td>
<td>Alu031</td>
<td>AF271434</td>
<td>microsatellite</td>
</tr>
<tr>
<td>9</td>
<td>Alu032</td>
<td>AF271435</td>
<td>microsatellite</td>
</tr>
<tr>
<td>10</td>
<td>Alu034</td>
<td>AF271436</td>
<td>microsatellite</td>
</tr>
<tr>
<td>11</td>
<td>Alu038</td>
<td>AF271439</td>
<td>microsatellite</td>
</tr>
<tr>
<td>12</td>
<td>Alu039</td>
<td>AF271440</td>
<td>microsatellite</td>
</tr>
<tr>
<td>13</td>
<td>Alu046</td>
<td>AF271443</td>
<td>microsatellite</td>
</tr>
<tr>
<td>14</td>
<td>Alu048</td>
<td>AF271444</td>
<td>microsatellite</td>
</tr>
<tr>
<td>15</td>
<td>Alu054</td>
<td>AF271446</td>
<td>microsatellite</td>
</tr>
<tr>
<td>16</td>
<td>Alu059</td>
<td>AF271448</td>
<td>microsatellite</td>
</tr>
<tr>
<td>17</td>
<td>Alu062</td>
<td>AF271449</td>
<td>microsatellite</td>
</tr>
<tr>
<td>18</td>
<td>Alu070</td>
<td>AF271451</td>
<td>microsatellite</td>
</tr>
<tr>
<td>19</td>
<td>Alu072</td>
<td>AF271452</td>
<td>microsatellite</td>
</tr>
<tr>
<td>20</td>
<td>Alu073</td>
<td>AF271453</td>
<td>microsatellite</td>
</tr>
<tr>
<td>21</td>
<td>Alu1</td>
<td>AF271465</td>
<td>microsatellite</td>
</tr>
</tbody>
</table>

(*) For duplicated loci we have used /1 or /2 for the two loci. For markers derived from the same clone (sequence) we have used A and B and for some also C (were we designed three separate markers from the same clone).
MapView

All markers “clickable”
Marker view

Genetic Map, markers

Description for marker BHMS7-029

- **Microsatellite name:** Ssa0004NVH
- **Organism:** Salmon salar, Atlantic salmon
- **Clone name:** BHMS7-029
- **GenBank accession:** AF256647
- **Forward primer:** AGGTACCCTGCTTTTCAC
- **Reverse primer:** ATGTAAGGAGCCATAAAG
- **Low size:** 223
- **Upper size:**
- **Comments:**
- **Temperature:** 52.0°C
- **MgCl:** 1.5 mM

Return to all markers
Salmon Genome Project (SGP)

- Genetic map
- BAC library/physical map
- EST-sequencing
- Expression profiling
- QTL studies
- Genomic sequencing
- Bioinformatics

www.salmongenome.no
**BAC library**

**BAC library construction:**

- Constructed by Jim Thorsen
- Collaboration with Pieter de Jong (USA) and the Canadian Genomic Research on Atlantic salmon Project (GRASP)

  - 313,000 clones picked and gridded
  - average insert size 190kb
  - CHORI-214 (~ 18 x coverage)
- Copies in Norway and Canada
- Main distribution Dr. Pieter de Jong
  bacpac.chori.org
- High density filters

**Size distribution of insert**

Thorsen et al 2005 BMC Genomics 6:50
Fingerprinting of BAC DNA

- Most of the fingerprinting done in Canada
  - Genome Sciences Centre (BCCA)
- 186 000 clones HindIII fingerprinted
- 37 285 BACs as singletons
- 4 354 contigs

Ng et al 2005 Genomics 86: 396-404
Salmon Genome Project (SGP)

- Genetic map
- BAC library/physical map
- EST-sequencing
- Expression profiling
- QTL studies
- Genomic sequencing
- Bioinformatics

www.salmongenome.no
Sequencing pipeline

- Sequencing
- Verifying results / “cleaning up”
  - removing bad sequence
- Clustering
  - how many are from the same gene?
- BLAST/Annotation
  - Which genes do we have?
cDNA-libraries

- Developed tissue and developmental specific libraries
  15 tissues from pre- and post-smolt

Tissues
- Spleen
- Kidney
  Headkidney
  Kidney
- Heart
- Brain
- Swimbladder
- Skin
- Muscle
  White
  Red
- Liver
- Ovaries
- Testes
- Eye
- Gills
- Intestine

- All directionally cloned
- Picked and gridded primary clones
- Approx. 11 500 clones from each tissue
Sequencing

- Sequenced approx. 68,000 ESTs from pre-smolt - Primarily 5’ sequences
- Sequenced over 1,100 full length cDNAs
- Developed a “preAssemble” pipeline for automatic sequence processing
  - A chance to take out bad quality sequences prior to clustering
Clustering

- Clustered 56,000 sequences
  - 6,203 contigs
  - 13,220 singlets
  - 19,423 total

- Currently clustering and annotating all salmon ESTs available
  - 189,000 (will be 300,000 in near future)

- Also doing tissue specific clustering

- Scripts for pipelines to automatic clustering, BLAST against pdb, swissprot, nr and nt and annotation based on GO with outputs in html format
SNP identification pipeline

- Identification of putative SNPs by running Phred – Phrap – Poly Bayes software on the data produced by sequencing machines (raw sequence data). This data is stored in the SGP database.

- Results produced by the Pre-assembly sequence processing pipeline are used to create a “good” data subset
Annotation

- Automatic annotation based on Gene Ontology (GO) using swissprot
- Organised according to:
  - Molecular function
  - Biological process
  - Cellular component
Gene Ontology (GO) Molecular Function annotation for sequences:

Gill and intestine cDNA library EST contigs/singletons have been automatically annotated by applying BLAST alignment results (blastx against the swissprot database) generated with our standalone version of NCBI BLAST. Sequences with BLAST hits with E-value lower than 1.0e-15 have been annotated by applying the GO assignments for the UniProt database produced by the GOA project of the European Bioinformatics Institute. The gene_association goa_uniprot database of 26.04.2004 was used (See the EBI site for details) together with GO terms from the GO release of 11.05.2004 (See the GO site for details). The sequences have been annotated based on the single best hit in the swissprot database.

Parent-child relationships are represented by indentation in the table below. Note that for any GO child term that describes a gene product, all its parent terms also apply to that gene product. This rule has been used when constructing the table below. Follow the links below to reach the sequences assigned to each GO term (or to one of its children terms). A "part of" relationship is denoted by "<", while "%" is used for "is a" relationships.

| %antioxidant activity: GO:0016209 | 0.00158 | 0.00163 | 0.00712 | 0.00761 | 0.00764 | 0.00971 | 0.01353 | 0.01278 | 0.02569 | 0.02574 |
| %glutathione-dissulfide reductase activity: GO:004262 | 0.00097 | 0.01324 |
| %peroxidase activity: GO:0004601, GO:0016685, GO:0016686, GO:0016687, GO:0016688 | 0.00163 | 0.00761 | 0.00764 | 0.02569 | 0.02574 |
| %binding: GO:0005488 | 0.00002 | 0.00005 | 0.00009 | 0.00015 | 0.00017 | 0.00020 | 0.00025 | 0.00036 | 0.00041 | 0.00044 | 0.00046 |
| 0.00059 | 0.00061 | 0.00062 | 0.00066 | 0.00069 | 0.00071 | 0.00077 | 0.00078 | 0.00079 | 0.00080 | 0.00081 | 0.00083 |
| 0.00088 | 0.00090 | 0.00092 | 0.00098 | 0.00100 | 0.00104 | 0.00105 | 0.00108 | 0.00109 | 0.00110 | 0.00111 | 0.00112 |
| 0.00127 | 0.00128 | 0.00130 | 0.00133 | 0.00135 | 0.00145 | 0.00148 | 0.00153 | 0.00153 | 0.00163 | 0.00167 | 0.00178 | 0.00179 |
| 0.00181 | 0.00182 | 0.00184 | 0.00185 | 0.00188 | 0.00191 | 0.00193 | 0.00195 | 0.00200 | 0.00202 | 0.00203 | 0.00207 |
| 0.00229 | 0.00221 | 0.00222 | 0.00226 | 0.00227 | 0.00232 | 0.00233 | 0.00236 | 0.00240 | 0.00249 | 0.00249 | 0.00249 |
| 0.00251 | 0.00253 | 0.00262 | 0.00267 | 0.00274 | 0.00277 | 0.00285 | 0.00292 | 0.00301 | 0.00318 | 0.00324 | 0.00324 |
| 0.00326 | 0.00328 | 0.00343 | 0.00358 | 0.00365 | 0.00369 | 0.00375 | 0.00379 | 0.00381 | 0.00382 | 0.00382 | 0.00392 |
| 0.00421 | 0.00442 | 0.00443 | 0.00445 | 0.00450 | 0.00452 | 0.00453 | 0.00456 | 0.00456 | 0.00457 | 0.00458 | 0.00458 |
| 0.00450 | 0.00452 | 0.00453 | 0.00458 | 0.00480 | 0.00483 | 0.00485 | 0.00489 | 0.00490 | 0.00491 | 0.00494 | 0.00495 | 0.00500 |
| 0.00509 | 0.00510 | 0.00511 | 0.00513 | 0.00516 | 0.00518 | 0.00522 | 0.00523 | 0.00525 | 0.00526 | 0.00526 | 0.00534 | 0.00536 |
| 0.00537 | 0.00540 | 0.00541 | 0.00542 | 0.00544 | 0.00549 | 0.00551 | 0.00552 | 0.00553 | 0.00556 | 0.00563 | 0.00567 | 0.00570 |
| 0.00575 | 0.00578 | 0.00588 | 0.00595 | 0.00600 | 0.00610 | 0.00613 | 0.00617 | 0.00622 | 0.00624 | 0.00626 | 0.00629 |
| 0.00627 | 0.00642 | 0.00644 | 0.00646 | 0.00649 | 0.00659 | 0.00662 | 0.00665 | 0.00671 | 0.00673 | 0.00674 | 0.00677 |
Salmon Genome Project (SGP)

- Genetic map
- BAC library/physical map
- EST-sequencing
- Expression profiling
- QTL studies
- Genomic sequencing
- Bioinformatics

www.salmongenome.no
SGP computational resources

- **Linux workstation**
- **Linux workstation**
- **Windows workstation**
- **Sun server**
  - Disk space 420 GB
- **Oracle database**
- **SGP web server**
- **HP Super Dome**
  - SGP: 4 processors
  - Disk space 10x73 GB
  - BLAST and pipelines
- **HP Itanium server**
  - 4 CPU
  - Disk space 2500 GB
- **EMBL database**
- **GenBank database**
- **Batch jobs**
- **Data transfer**
SGPweb resources

- Web site
- Database (Oracle)
- Query system and interface
- Data loading
- Sequence processing and analysis
- User access
- Project Information
- BLAST service

www.salmongenome.no
SGP Blast service

- **Public blast**
- **Internal blast**
  - Up to 100 sequences
  - Up to 36 hours on the HP Super Dome supercomputer
  - Results can be received by email
  - Results can be saved as HTML
  - Run blast directly from the database query results page
All EMBL databases

All NCBI polynucleotide databases

All NCBI polypeptide databases

Local SGP database (Currently only "Private BLAST")
Salmon Genome Project (SGP)

- Genetic map
- BAC library/physical map
- EST-sequencing
- Expression profiling
- QTL studies
- Genomic sequencing
- Bioinformatics
Microarray
Atlantic salmon

Canadian cDNA chip
- Atlantic salmon and rainbow trout sequences

Progress:
- 3 700 gene preliminary Canadian chip
- Tested in Canada and Norway

Tested for:
- Cross-species
- Various disease challenged fish
- Smoltification
- Environmental issues (polluted rivers)

Next generation:
- 16 000 gene chip (Canadian)
- June 2004

A second chip has been developed as a collaboration between the UK Salmon Traits Project and SGP
- 17 500 genes in duplicate
# Acknowledgement

**SGP:**
- Bjørn Høyheim
- Heidi Hagen-Larsen
- Jim Thorsen
- Sigbjørn Lunner
- Ole Albert Guttersrud
- Kristin Vekterud
- Jill Andersen
- Agate Hansen
- Katrine Hånes
- Alexei Adzhubei
- Anna Sadym
- Jon Lærdahl
- Rune Male
- Sigve Ladstein
- Hee-Chan Seo
- Sutada Mungpakdee
- Knut Jørstad
- Inge Jonassen
- Bjarte Dysvik
- Ketil Malde
- Finn Drabløs
- Heather Wieman

**SGP**
- Former coworkers

**GRASP:**
- William S. Davidson
- Ben F. Koop

**GRASP**
- Christian Bendixen
- Lars-Erik Holm
- Frank Panitz

**SALMAP:**
- Bjørn Høyheim
- Monica Skogen
- Sven Martin Jørgensen
- John Taggart
- Margaret Cairney

**SALMAP**
- Richard Powell
- Grace Davey

**SALGENE:**
- Bjørn Høyheim
- Heidi Hagen-Larsen
- Agate Hansen
- Richard Powell
- Grace Davey

**SALGENE**
- John Taggart
- Margaret Cairney

**CHORI:**
- Pieter de Jong
- Dr. Kazutoyo Osoegawa
- Dr. Baoli Zhu
- Gery Vessere

**CHORI**
- Richard Powell
- Gavin Cloherty
- René Guyomard
- Karim Gharbi

**CHORI**
- Roy Danzmann
- Moira Fergusson
- Nobuaki Okamoto
- Takashi Sakamoto

**CHORI**
- United Gene Institute
- Shanghai
- Ying Kang
- Co-workers

**TRAITS (UK)**
- Alan Teale
- John Taggart
- Chris Secombes
- Sam Martin
- Glen Sweeney
- Paul Dear

**TRAITS (UK)**
- United Gene Institute
- Shanghai
- Ying Kang
- Co-workers