Organization and synteny of hemoglobin gene cluster in codfishes: our experience with targeted sequence capture

Ave Tooming-Klunderud, NSC and CEES at UoOslo,
ave.tooming-klunderud@ibv.uio.no
www.sequencing.uio.no
Leiden, 02/05/2017
Norwegian Sequencing Centre

• Two nodes:
  • CEES at Department of Biosciences, University of Oslo
  • Department of Medical Genetics, Oslo University Hospital

• Funding:
  • Host institutions
  • Research Council of Norway
    • NSC - INFRAstructure program 2009-2013
    • NSC Phase II – INFRAstructure 2014 – 2017
    • NorSeq (National consortium for sequencing and personalized medicine) – INFRAstructure 2015 – 2018
Research of Atlantic cod at NSC

LETTER

The genome sequence of Atlantic cod reveals a unique immune system

Bastiaan Star1, Alexander J. Nederbragt1, Sissel Jentoft1, Unni Grieg1

www.nature.com/scientificreports

OPEN Genomic characterization of the Atlantic cod sex-locus

Torresen et al. BMC Genomics. (2017) 18:95
DOI 10.1186/s12864-016-3448-x

Unraveling the Evolution of the Atlantic
cod genome

RESEARCH ARTICLE

An improved genome assembly unravels prolific tandem repeats in Atlantic cod

Ole K. Torresen1, Bastiaan Star1, Sissel Jentoft1, William B. Reinar1, Harald Groe2, Brian P. Walen2, James Knight2, Jenny M. Ekholm2, Paul Peluso2, Rolf B. Edvard Ave Tooming-Klunderud2, Morten Skage1, Sigbjørn Lien3, Kjetill S. Jakobsen1 and Alexander J. Nederbragt1

www.nature.com/scientificdata

Data Descriptor: Whole genome sequencing data and de novo draft assemblies for 66 teleost species

Martin Malmström3, Michael Matschiner2, Ole K. Torresen2, Kjetill S. Jakobsen1 & Sissel Jentoft1,2
EVOLUTION OF HEMOGLOBIN GENES IN CODFISHES
Adaption to different environments

• Hemoglobins ability to bind O_2 is influenced by environmental factors
  • Temperature
  • Hydrostatic pressure
• Temperature adaptation of hemoglobin in Atlantic Cod - polymorphism in Hb gene β1

β1 locus
β1-1 (Met55Lys62)
β1-2 (Val55Ala62)

Hemoglobin β1-1 frequencies of cod in the North Atlantic. Ross et al. 2013
Hemoglobin genes in fish

Hb cluster in the common ancestor of teleost fish and gar

Hb cluster in the pre TGD teleost ancestor

TGD

Hb clusters in the post TGD teleost ancestor

Opazo et al. 2012
Hemoglobin genes in codfishes

What about synteny?
Two alternatives:

- Generate high quality reference genomes
- Sequence a subset of the genome by targeted sequence capture
Custom probe design – Roche Nimblegen

- Sequences used for the probe design
  - The target region (MN and LA clusters) in Atlantic cod reference genome

- Sequences of Hemoglobin and flanking genes
- Partially assembled genomes for additional species – only regions orthologous to the target region of Atlantic cod
Phylogenetic relationship of the eight species of codfishes included in the study. Time since divergence given at each node in millions of years. Subset from Malmstrøm et al. 2016
Pre capture multiplexing

• 16 bp PacBio barcode

SeqCap Oligo1: AATGATACGGGCAACCACCCGAGA
PreCap-2 fwd: GTCAGACGATGCGTCAT AATGATACGGGCAACCACCCGAGA……
Block-2 fwd: GTCAGACGATGCGTCAT ------HE blocking from kit--------
PostCap-2 fwd: GTCAGACGATGCGTCAT AATGA

PreCap-2 rev: CGCGATCTATGCACACCGCAAGCAGAAGACGGCATAACGAG…..
SeqCap Oligo2 : CAAGCAGAAGACGGCATAACGAG
Block-2 rev: CGCGATCTATGCACACCG ------HE blocking from kit--------
PostCap-2 rev: CGCGATCTATGCACACCGCAAGC

• 6 bp Illumina barcode
SeqCap EZ Enrichment System

Workflow for long fragments

1. Target Regions
2. Prepare with Next-Gen Sequencing Adaptors
   - Covaris shear
   - KAPA library prep kit
3. SeqCap EZ Probes (Solution Capture)
   - Dual SPRI size selection then amp with KAPA HiFi
4. Amplify DNA and Enrichment QC
   - Post-capture amplification with KAPA HiFi
5. Sequence DNA on a Next-Gen Sequencer

Genomic DNA → Library Preparation → Hybridization → Capture and Washing → Amplification and Sequencing

- Genomic DNA
- Library Preparation
- Hybridization
- Capture and Washing
- Amplification and Sequencing

- G-tube shear
- KAPA Hyper Prep
- Long range PCR with Takara enzyme
- Capture with Invitrogen M-270 beads
- Post-capture long range PCR with Takara enzyme
Megaruptor
BluePippin
used for one sample
PCR with KAPA HiFi enzyme, all samples barcoded with PacBio barcodes during PCR

Modifications of the protocol
Library prep and capture

- 4 samples pooled before capture
- Individual captures for 6 of the samples, similarly sized post capture PCR products pooled into two final libraries
- All 8 samples pooled before capture, 250 ng of each
Sequencing results

Pool of 8 samples

% of reads

<table>
<thead>
<tr>
<th>Species</th>
<th>% of reads</th>
<th>#reads</th>
<th>#bases</th>
<th>av. length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic cod</td>
<td>12.6%</td>
<td>84304</td>
<td>258 Mb</td>
<td>3.06 kb</td>
</tr>
<tr>
<td>Haddock</td>
<td>5.4%</td>
<td>42530</td>
<td>132 Mb</td>
<td>3.1 kb</td>
</tr>
<tr>
<td>Silvery pout</td>
<td>17.1%</td>
<td>86067</td>
<td>268 Mb</td>
<td>3.1 kb</td>
</tr>
<tr>
<td>Cusk</td>
<td>12.2%</td>
<td>66450</td>
<td>214 Mb</td>
<td>3.2 kb</td>
</tr>
<tr>
<td>Burbot</td>
<td>7.1%</td>
<td>64472</td>
<td>195 Mb</td>
<td>3.01 kb</td>
</tr>
<tr>
<td>European hake</td>
<td>21.5%</td>
<td>76604</td>
<td>215 Mb</td>
<td>2.8 kb</td>
</tr>
<tr>
<td>Marbled moray cod</td>
<td>9%</td>
<td>64202</td>
<td>186 Mb</td>
<td>2.9 kb</td>
</tr>
<tr>
<td>Roughhead grenadier</td>
<td>15.2%</td>
<td>54391</td>
<td>155 Mb</td>
<td>2.85 kb</td>
</tr>
</tbody>
</table>
«Jumping» PacBio barcodes

Due to our barcode design:

SeqCap Oligo1: AATGATACGGCGACCACCAGAG
PreCap-2 fwd: GTCAGACGATGCGCTCAT AATGATACGGCGACCACCAGAG......
Block-2 fwd: GTCAGACGATGCGCTCAT       ------HE blocking from kit---------
PostCap-2 fwd: GTCAGACGATGCGCTCATAATGA

PreCap-2 rev: CGCGATCTATGCACACGCAAGCAGAAGACGGCATACGAG.....
SeqCap Oligo2: CAAGCAGAAGACGGCATACGAG
Block-2 rev: CGCGATCTATGCACACG       ------HE blocking from kit---------
PostCap-2 rev: CGCGATCTATGCACACGCAAGC

some degree of barcode «jumping» was expected
«Jumping» PacBio barcodes

| 1486 | CGCAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTTCAGACG |
| 84   | GCAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTTCAGACG |
| 57   | ACGCAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTTCAGAC |
| 51   | CGCAAGCAGAAGACGGCATACGAGATACGGAAACTGTGACTGGAGTTTCAGA |
| 29   | CGCAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTTCAGACGT |
| 28   | TGCAAGCAGAAGACGGCATACGAGATTTGGTCA GTGACTGGAGTTTCAGACG |
| 28   | TCCAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTTCAGACG |
| 23   | ACCAAGCAGAAGACGGCATACGAGATCGTTTCACGTGACTGGAGTTTCAGA |
| 22   | GCCAAGCAGAAGACGGCATACGAGATGATCTGGCTGGAGTTTCAGACG |
| 16   | TCCAAGCAGAAGACGGCATACGAGATCGGGACGGGTGACTGGAGTTTCAGA |
| 12   | CGCAAGCAGCAGAAGATGAGACTGGACGCACACACACACACACACACACACACAC |
| 11   | CTCAAGCAGAAGACGGCATACGAGATCAGTGTGACTGGAGTTTCAGACG |

9.8% reads with wrong adapter

% of reads with wrong adapter for other samples:
0.4%  0.8%  1.5%  2.4%  2.5%  3.4%  5.0%
Data analysis workflow

Raw reads (fastq) → Filtering & Adapter trimming → De novo Assembly → Assembly assessment → Analysis

Bioawk Qiime Prinseq → Mapping → Canu

BWA-mem → Quast

Raw reads = reads filtered and demultiplexed using RS_Reads of Insert.1 pipeline
## De novo assembly

<table>
<thead>
<tr>
<th></th>
<th>Total length (&gt;= 1000 bp)</th>
<th>Total number unitigs</th>
<th>Largest unitigs (bp)</th>
<th>GC (%)</th>
<th>N50</th>
<th>N75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic Cod</td>
<td>1757471</td>
<td>215</td>
<td>41343</td>
<td>44.91</td>
<td>8575</td>
<td>5750</td>
</tr>
<tr>
<td>Haddock</td>
<td>971706</td>
<td>116</td>
<td>44949</td>
<td>45.20</td>
<td>8776</td>
<td>6169</td>
</tr>
<tr>
<td>Cusk</td>
<td>2232327</td>
<td>248</td>
<td>117090</td>
<td>42.62</td>
<td>9260</td>
<td>6405</td>
</tr>
<tr>
<td>Silvery cod</td>
<td>1713941</td>
<td>210</td>
<td>41689</td>
<td>45.41</td>
<td>8649</td>
<td>5879</td>
</tr>
<tr>
<td>Burbot</td>
<td>943952</td>
<td>103</td>
<td>58914</td>
<td>44.23</td>
<td>10360</td>
<td>6386</td>
</tr>
<tr>
<td>European Hake</td>
<td>1410818</td>
<td>197</td>
<td>36393</td>
<td>46.01</td>
<td>7037</td>
<td>5400</td>
</tr>
<tr>
<td>Roughhead grenadier</td>
<td>1407647</td>
<td>182</td>
<td>42357</td>
<td>48.44</td>
<td>8192</td>
<td>5655</td>
</tr>
<tr>
<td>Marbled moray Cod</td>
<td>1552724</td>
<td>206</td>
<td>38697</td>
<td>45.02</td>
<td>7963</td>
<td>5539</td>
</tr>
</tbody>
</table>
Synteny of the hemoglobin clusters

MN cluster

LA cluster

Number of genes detected from draft assemblies indicated on tree
Evaluation of capture - Atlantic cod

- Probes do not cover entire sequence due to repeats

Dinucleotide AG (%)

[Dinucleotide AC (%)]

Star et al. 2016

Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples

Bastiaan Star, Marianne HS Hansen, Morten Skage, Ian R. Bradbury, Jane A. Godiksen, Olav S. Kjesbu, and Sissel Jentoft
Repeat content is 20.29% in LA and 10.7% in MN cluster.
Evaluation of capture – rest of the fishes

- Capture efficiency decreases with increased evolutionary distance from Atlantic cod
Acknowledgments

• Cod-group
  • Siv Nam Khang Hoff
  • Helle Baalsrud
  • William B. Reinar
  • Sissel Jentoft
  • Kjetill S. Jakobsen
  • Ole Kristian Tørresen
  • Martin Malmstrøm

• NSC
  • Morten Skage
  • Marianne H.S. Hansen
  • Lex Nederbragt

• Roche/Nimblegen
  • Reza Shirzadi
  • Gregor Obernosterer
  • Todd Richmond