**DNA Sequencing & Otolith Microchemistry Data Set**

**Pacific cod, South Korea coastal waters**

**Dataset Description**

This data set consists of genetic and otolith microchemistry data for Pacific cod, for the determination of Pacific cod stock structure around the Korean peninsula. Pacific cod spawn in aggregations during the winter months, with studies suggesting that they display site fidelity to their natal aggregation. South Korean fisheries managers are interested in the spatial scale at which spawning aggregates display significant genetic differentiation, and if that differentiation can be correlated with site fidelity to aggregates on ecological timescales.

Mature Pacific cod were sampled from winter spawning aggregates in coastal South Korean waters over the course of several spawning seasons. For each individual, a fin clip or tissue sample was taken and preserved in 100% denatured ethanol, and either one or both otoliths (ear bones) collected and stored dried. Samples were collected by Dr. Sukyung Kang of the National Institute of Fisheries Science (spawning seasons 2014 – 2016), and Dr. WooSeok Gwak of Gyeongsang National University (2007-2008 spawning season). Tissue samples provided by Dr. WooSeok Gwak are subsamples from the original samples taken in 2007-2008.

The sampled spawning aggregates span three regions of South Korean coastal waters: the western coast (2 sites), southern coast (3 sites), and eastern coast (1 site). One site is considered classified as “southeastern coast” because of its potential location to the south or east of a critically important winter current. The data also includes temporal replicates; one site was sampled during both the 2013-2014 and 2014-2015 spawning seasons to test between-year differences, and one site sampled at the beginning and end of the 2007-2008 spawning season to test within-year differences.

Genetic data collection consisted of restriction-site associated DNA sequencing, which produced thousands of single nucleotide polymorphisms (referred to as “markers”). DNA sequencing preparation was conducted at the University of Washington, and sequencing completed at the University of Oregon’s Genomics & Cell Characterization Core Facility. Otoliths were processed and aged by the Alaska Fisheries Science Center’s Age & Growth Program. Otolith microchemical data collection was completed using the University of Oregon’s mass spectrometer.

After data filtering and processing, genetic data is available for 253 individuals. Of the 253 individuals processed for genetic analysis, 198 individuals have corresponding otolith data.

**Initial Thoughts**

The genetic data set can determine the population structure of Pacific cod around the Korean peninsula on evolutionary timescales, by conducting multivariate analyses on genetic markers to group individuals into distinct clusters (assume this will be explained by sampling site, spawning season, and/or region). I am primarily interested in: (1) Is there significant differentiation between spawning aggregates between regions? And (2) Is there significant differentiation within regions?

The purpose of the otolith microchemistry analysis is to determine population structure on ecological timescales. To do so, multivariate analyses will be conducted on otolith microchemical concentrations of the bone layer at (1) the core of the otolith, representative of the body of water where the fish spent time as an early juvenile, and (2) the edge of the otolith, representative of the body of water where the fish was caught. This comparison will show the extent to which fish show site fidelity to their natal spawning population. The comparison is dependent on different aggregates spawning in bodies of water with different chemical compositions; luckily, initial analyses show distinctive microchemical signatures for certain spawning aggregates.

I would like to figure out a way to combine the two datasets in an exploratory multivariate analysis of how population structure differs, or remains the same, on evolutionary versus ecological timescales – primarily through a PCA graph or similar visualization. At the moment I think that the best way to do this would be to conduct separate PCAs on the genetic and otolith data and then use the first principal component from each to plot all of the data on the same graph. Another option might be to use the otolith microchemistry as an explanatory variable, with genetic markers as the response variable.

**Other data**

It would be interesting to access existing environmental data, particularly water temperature data, for the dates and locations where samples were taken. Pacific cod spawn at very specific temperatures, with different aggregates spawning at different water temperatures (on a latitudinal cline). Initial genetic analyses also show that certain spawning aggregates have an unexpectedly low effective population size; with biomass data for these spawning aggregates, we could see if recent population bottlenecks from overfishing could be causing this historical signature.

Neither of these data sets are readily available or translated to English; I would contact the National Institute of Fisheries Science to see if they have a similar buoy / biomass survey system as NOAA.

**Metadata Tables**

Response Variables

1. Genetic markers
   1. 5,405 single nucleotide polymorphisms (SNPs), recorded as “Locus\_” (too many to put in a table)
2. Otolith chemical markers

|  |  |  |  |
| --- | --- | --- | --- |
| **Chemical Elements** | | | |
| **Abbr.** | **Element** | **Units** | **Range** |
| B11 | Boron | Ci:CCa | 0.63 – 1.84 |
| Ba138 | Barium | Ci:CCa | 1.28 – 2.96 |
| Li7 | Lithium | Ci:CCa | -0.18 – 0.81 |
| Mg25 | Magnesium | Ci:CCa | 3.85 – 5.96 |
| Mn55 | Manganese | Ci:CCa | 0.24 – 1.41 |
| Pb208 | Lead | Ci:CCa | 0.01 – 1.41 |
| Si86 | Silicon | Ci:CCa | 8.07 – 9.19 |
| Zn66 | Zinc | Ci:CCa | 0.78 – 3.28 |

Explanatory Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Abbr.** | **Full Name** | | **Units** |
| **Sampling Site** | | | |
| BOR | | Boryeong |  |
| YS | | Yellow Sea Block 161 |  |
| NAM | | Namhae |  |
| GE / GEO | | Geoje |  |
| JB | | Jinhae Bay |  |
| PO | | Pohang |  |
| BOR | | Boryeong |  |
| **Spawning Season** | | |  |
| ‘07 | | Winter 2007-2008 |  |
| ‘14 | | Winter 2013 - 2014 |  |
| ‘15 | | Winter 2014 – 2015 |  |
| ‘16 | | Winter 2015 - 2016 |  |
| **Region** | | |  |
| W | | West Coast |  |
| S | | Southern Coast |  |
| SE | | Southeastern Coast |  |
| E | | Eastern Coast |  |
| **Meristics** | | |  |
| TL | | Total length of fish | Centimeters |
| W | | Weight of fish | Grams |
| GW | | Gonad weight of fish (g) | Grams |
| **Sex & Maturity** | | |  |
| Sex | | Sex of fish\* | M/F |
| Maturity | | Sexual maturity of fish at time of capture\* |  |
| Age | | Age of fish, as determined from otolith | Years |
| OW | | Otolith weight (g) | Grams |
| **Life stage of Otolith measurement** | | |  |
| J | | Juvenile (measurement from core of otolith) |  |
| A | | Adult (measurement from edge of otolith) |  |

\* Low confidence that these variables were correctly measured in the field. Will likely remove from final data set.

**Additional Notes**

Individual Codes

Each individual sample is coded according to sampling site, sampling date. Following the code is a specific numerical identifier for that sample, and any appropriate suffixes (see next section).

|  |  |  |  |
| --- | --- | --- | --- |
| **Abbreviation** | **Sampling Site Location** | **Sampling Date** | **Analysis Group** |
| PO010715 | Pohang | 1/7/2015 | PO ‘15 |
| PO020515 | Pohang | 2/5/2015 | PO ‘15 |
| GE011215 | Geoje | 1/12/15 | GE ‘15 |
| GE012315 | Geoje | 1/23/2015 | GE ‘15 |
| GEO020414 | Geoje | 2/4/2014 | GEO ‘14 |
| YS121316 | Yellow Sea Block 161 | 12/13/2016 | YS ‘16 |
| JB121807 | Jinhae Bay | 12/18/2007 | JB ‘07, Early Spawn |
| JB021108 | Jinhae Bay | 2/11/2008 | JB ‘07, Late Spawn |
| JUK07 | Jukbyeon | 12/10/2007 | JUK ‘07 |
| BOR07 | Boryeong | 1/24/2007 | BOR ‘07 |

Individual Suffixes

Describe genetic data only. Suffixes on individual codes provide an additional description of how the sample was processed. This was included in the data set because the difference in processing may affect data quality, and so these samples were tracked through data filtering steps.

|  |  |
| --- | --- |
| Suffix | Processing |
| “\_2” | The sample used in the data set is a re-run of the original sample. This may have been done because of contamination, low read depth, or low DNA quality. |
| “\_c” | The sample used in the data set is a combination of two separate DNA extractions, because of low quantities of DNA in either extraction. |
| “\_rep” | The sample used in the data set was originally prepared as a replicate for an existing sample. The replicate showed higher quality or read depth, and so was maintained in the data set instead of the original sample. |
| “.1” | The sample was put through paired end sequencing. Whereas only forward sequencing reads are available for samples without this prefix, “.1” samples have both forward and reverse reads available for data analysis. The reverse reads will not be used for this project. |