8. Data Management Plan

8.1 Biological and Chemical Data (Taken from Data Management Plan previously provided in a broader context to the OAP by the "NWFSC Sustained Investment Program")

Coverage: Biological and chemical data collected and/or created under this award will be made visible, accessible, and independently understandable to users in a timely manner, consistent with the requirements of NOAA's OAP.

Summary description of the data to be generated: Data collected for this project will be biological measurements made on geoduck larvae exposed to ambient and elevated CO2 levels and temperatures and descriptions of physical properties of seawater in the experimental treatments. Data will also include extensive sequence information that is covered separately below. The biological measurements will include survival, morphometrics, developmental stage, and presence or absence of calcification in larvae. Data on seawater characteristics will include temperature, salinity, pH, dissolved inorganic carbonate, total alkalinity, and CO2 concentration. This project will not include the collection of "environmental data" as defined as geospatially-referenced observations reflecting the natural biological, physical, or chemical conditions of the ocean or atmosphere.

Data types: Data will generally be comprised of digital numeric data.

Quality Assurance: Quality control on chemical measurements of MOAT environments will be performed as part of the analysis of samples by PMEL including instrument calibration, standards and appropriate blanks.

Availability: All biological and chemical data will be immediately archived and stored on the PI's local server with RAID redundancy. After the completion of the replicated experiments and appropriate QA/QC procedures are conducted to determine if further experiment or analyses need to be conducted, the data will be uploaded to publicly accessible data repositories including the University of Washington Research Works (https://digital.lib.washington.edu/researchworks/) and the NOAA National Oceanographic Data Center (http://www.nodc.noaa.gov/) where data will be permanently archived.

Responsible Parties: Rick Goetz (rick.goetz@noaa.gov) will be the primary point of contact for general questions about this project; Paul McElhany (paul.mcelhany@noaa.gov) concerning questions related to MOATS and their operation; and Steven Roberts (sr320@uw.edu) concerning questions related to sequencing and genomic analyses. NODC will be responsible for permanent data archiving and ensuring public availability of the data.

8.2 Sequence Data

Data Description: As part of this research effort a significant amount of sequence data will be generated including PacBio long-read sequencing data (genomic scaffolding) and short read sequencing information from the Illumina Hi-Seq platform (RAD-Seq and BS-Seq). A majority of the sequencing data will be from the Illumina platform and will include files in fastq format where both quality and nucleotide information is included. These files will be the basis of analysis that will primarily be carried out with bisulfite sequence alignment software (BSMAP), SNP characterization software (STACKS), and downstream statistical packages (R, python). This project will generate approximately 15 lanes of sequence data with each lane resulting in a file size (compressed) of approximately 11 gigabytes / lane.

Access and Sharing: Raw data from DNA sequence platforms will be transferred to the lab of a collaborator (Roberts) where an internal sequence data management plan is followed (https://github.com/sr320/LabDocs/blob/master/DMPseq.md). Raw data is organized on a network attached storage (NAS) device with RAID redundancy. This NAS is open to the public (http://owl.fish.washington.edu/nightingales/). To make it easier for searching and discovery we also maintain a separate database including metadata (see below) and direct links to files (http://goo.gl/XxjTkW). Within one month of acquiring raw data it will be uploaded into the NCBI Short Read Archive (SRA) database. All data will be released once the results are published or no later than three months after the project end date. Raw data from secondary procedures including mapping and genome feature analysis will be available in real-time on the Roberts Lab wiki where all lab notebooks are open to the public (genefish.wikispaces.com). Data will be in non-proprietary formats such as tab-delimited text files. Limited analyzed data and workflows will also be made available via Galaxy and iPlant Collaborative as some analysis will take place on these platforms.

Metadata: The Roberts database includes information such as file name, data, taxa, tissue, molecule, platform, length, description, and file locations. Essential information including a description of the sample, library, and sequencing method will be included in the SRA repository. Data tags will allow the data to be easily retrievable at NCBI.

Archiving and Backup: As described above, raw data from DNA sequencing is stored on a NAS RAID server and will be uploaded to NCBI for archiving as well as providing access. Raw data will also be stored in two physical locations in the School of Aquatic Fishery Sciences at the University of Washington. Furthermore, all data located on http://owl.fish.washington.edu/nightingales is mirrored on the Amazon Glacier service. All lab notebooks where analysis is carried out are published via RSS feed and backed up to PDF. Notebooks in the Wordpress platform are archived using the XML export option on a regular basis.