Alan Wilson Centre Summer Scholarships 2014/2015 - RM13800 1246

Final Report – Jonathan Frericks, Superviser by Peter Ritchie. 17 March 2015

Hoki (*Macruronus novaezelandiae*) is New Zealand's most important commercial fish species, and represents the country's third most lucrative seafood product. It is a marine species that is phenotypically variable, has large populations and high gene flow, and occurs across an extensive and heterogeneous environment. However, relatively few studies have looked at hoki stock discrimination and population structure. Assembling nuclear and mitochondrial genomes will assist in the identification of nucleotide markers that can be used to investigate the population structure and gene flow of this economically important fish species. The positive outcomes from this project are diverse, and encompass several of the key objectives and missions of the AWC.

In July 2010, a *M. novaezelandiae* genomic DNA library was prepared and run in 2 lanes of an Illumina™ GAIIx. The average insert size was ~350bp with 66bp of adapter on each side. With a read length of 75 base (x2 for paired ends), this will generate an average distance between the two reads of ~200 bp. Sequencing was carried out using CASAVA Real Time Analysis 1.6.

The goal of this project was to analyse the sequence data, build contigs of the genome sequence and search for gene sequence matches in GenBank.

Before any analysis was done, FastQC was used to look at the quality and usability of the raw data – mostly for the purposes of comparison. To prepare the data for analysis, the adapters were removed and sequences of a minimum length of 36bp were retained. This was achieved using trimmomatic v0.3 and by obtaining the appropriate generation of Illumina adapter sequences. The resulting sequences were then analysed in FastQC again, demonstrating that out of ~28 million original fragments, ~17 million were retained for further analysis (60% retained).

Due to computational limitations, it was decided that the mitochondrial DNA would become the focus of the analysis using Geneious v8. The sequence data was aligned to a mitochondrial genome of a closely related species, European hake (*Merluccius merluccius*), accepting only 'perfectly paired' read mappings. Coverage was very good, except for the entire mitochondrial control region. This is unsurprising due to its hyper-variability. This incomplete *M. novaezelandiae* mitochondrial genome was then used as the reference genome and the process repeated. Areas with little or no coverage were then selectively sequenced with targeted primer pairs. Thus a complete mitochondrial genome was obtained.

A similar approach was taken for the nuclear analysis with Atlantic cod (*Gadus morhua*) as a reference nuclear genome using Burrows-Wheeler Aligner v 7.5a-r405. Approximately 2.9 million of the sequences mapping to the *G.morhua* genome and coverage was poor. Many of the sequences matched the same region of the *G. morhua* genome leaving large regions unmapped.

During this project an opportunity arose to investigate the use of MinION in a newly established laboratory. A 42ng sample of fragmented *M. novaezelandiae* DNA was prepared and Run on the MinION sequencer. The data from this run has not been analysed further.

This project has been highly beneficial for understanding the pipeline for genome sequencing, become familiar with the pertinent command line tools and even setting up the Linux environment in which these tools are run. I am interested in investigating this data further, beyond the scope of this summer studentship.

Further analysis on the *M. novaezelandiae* Illumina and MinION data should be done. Most significantly, de novo genome assembly should be a priority. This may be possible with Geneious v8 with greater RAM or may also be achieved with SOAP, Abyss, and Velvet on the Linux platform. The sequence data may also be useful for microsatellite and single nucleotide polymorphism discovery.

A peripheral use of the data may be as a teaching tool for students learning about genome assembly. Students will have access to computer laboratories with the analysis tools on it. Though the entire data set is large, it may be reduced so that students may easily be able to assemble or align the data themselves, allowing them to become familiar with the process.

I am extremely grateful to the AWC for this learning opportunity that has increased capability within the VUW as well as the AWC.