Scientific report: Assessing the impacts of inbreeding, population size and selection using RAD tag sequencing

4 months in the lab of Prof. Schlötterer in Vienna

1.1 Purpose of the visit

The main aim of my visit was to collaborate with skilled bioinformaticians and experienced researchers on a large dataset of raw genomic DNA obtained from temporal samples of 42 *Drosophila melanogaster* lines with Rad-Tag sequencing. Through this collaboration I hoped to increase my programming skills for bioinformatics, to build a pipeline for the analysis of pooled Rad-Tag data and process my data in this pipeline. Completing the above purposes would give me the qualifications to solve new bioinformatics challenges with this data and other data of value in the field of conservation genomics. In the last part of the exchange, the processing of the raw data should produce the preliminary results of the project at hand. These results are aimed at providing a greater insight into the effect of inbreeding, selection and their interaction across the genome.

1.2 Description of the work carried out during the visit

The raw Rad-Tag sequence data from my study was filtered and trimmed, leaving high quality sequences. These sequences was mapped to the *Drosophila melanogaster* reference genome with BWA (version 0.5.8c; Li & Durbin 2009) and GATK (version 2.4-7; McKenna *et al.* 2010). The mapping of the reads was validated by comparison to the expected sequencing pattern, and restriction sites with segregating sites removed. Indels and repeats were masked out with

PoPoolation2 (Kofler *et al.* 2011b). Tajima's pi, Watterson's theta and Tajima's D were estimated across windows and per position with PoPoolation (Futschik & Schlötterer 2010; Kofler *et al.* 2011a). SNP's under selection was identified with the CMH test implemented in PoPoolation, and the False Discovery Rate was calculated with an in house python script.

1.3 Description of the main results obtained

The successful processing of this large data set is a main result, as the final data is of very high quality and thus will be the starting point for many future analyses. During the exchange a few of these analyses were started, and different estimators of the genetic variation was calculated per position, with a sliding window and averaged across the genome. With these estimators a visualization and quantification of the genomic variation is possible, not only of the loss of genetic variation in different lines across time, but also how this loss of variation is different across the genome. SNP's under selection was also identified, but further work is needed before any conclusions can be made.

1.4 Future collaboration with host institution

A collaboration will continue to finish the analysis of data of this project, and other future collaborations is in my interest and highly plausible.

1.5 Projected publications

The data analyses are not yet completed, so the number of publications that can be produced from the work on the raw data performed in Vienna is not yet certain. At present time two publications are projected from this project; 1) A

paper with the focus on the adaptive potential of *D. melanogaster* to increasing temperatures, with the running title *"A laboratory natural selection experiment points to low evolutionary potential under natural temperature regimes in Drosophila melanogaster"* and 2) A paper with emphasis on the genomic effects of reduced effective population sizes, with the running title: *"Rad-Tag sequencing of Drosophila melanogaster exposed to increasing ambient temperatures and different levels of drift"*

1.6 Other comments (if any)

I gained experience with a number of different bioinformatic programs, improved my intuition for programming with both python and shell scripting, and also increased my molecular population genetics understanding. All abilities are important in the future work on my data. Furthermore I was inspired by the all the helpful researchers who aided me whenever I was stuck with a script or a problem of more conceptual understanding.

References

Futschik A. & Schlötterer C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, 186, 207-218.

Kofler R., Orozco-terWengel P., De Maio N., Pandey R.V., Nolte V., Futschik A., Kosiol C. & Schlötterer C. (2011a). PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *Plos One*, 6, e15925.

Kofler R., Pandey R.V. & Schlötterer C. (2011b). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, 27, 3435-3436.

Li H. & Durbin R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25, 1754-1760.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M. & DePristo M.A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20, 1297-1303.