Research Networking Programmes

Exchange Visit Grant

Scientific Report

*Application Reference N°*: 4572

Assessing the demographic history of Old World camelids (*Camelus* sp.) through whole genome sequencing

**Awardee:** Robert Fitak; postdoctoral researcher at the Veterinärmedizinische Universität Wien (*Vetmeduni* Vienna), Vienna, Austria.

**PI:** Pamela Burger; Group Leader at the Veterinärmedizinische Universität Wien (*Vetmeduni* Vienna), Vienna, Austria.

**In collaboration with:** Professor Michael W. Bruford and Pablo Orozco-terWengel of Cardiff University School of Biosciences, Cardiff, UK.

**Purpose of the visit**

The objective of this study was to apply two recently developed methods that infer the demographic history of populations from genetic data to a dataset of Old World camels (*Camelus* sp). This study was an initial component of a larger research program by our group to understand the evolutionary history of Old World camels, specifically to identify and conserve camel genomic diversity from both an agricultural and wildlife perspective. Each perspective requires characterization of both neutral levels of genomic variation in addition to loci influenced by selection, such as those responsible for local adaption and/or economically relevant traits. Unfortunately, locally reduced levels of genetic variation which are often a signature of selection can, at times, be indistinguishable from the effects of certain demographic processes (e.g. bottlenecks, inbreeding, etc). Despite the dogma of the selection vs. demography dilemma (‘selection acts on relatively small regions whereas demography affects the entire genome’), the stochastic nature of the coalescent can still lead to inconsistent patterns among loci. Nonetheless, understanding and accounting for the demographic history of a population is imperative when concluding which loci may be under the coercion of selection (Rosenberg and Nordborg 2002).

In addition to inferring the demographic history of our various camel populations, we found it necessary to investigate the effects of two important technical concerns often encountered when working with whole-genome resequencing data. First, does the level of divergence between an individual’s sequenced genome and the reference genome that was used for mapping reads have a significant impact on the observed demographic history? And second, does the removal of polymorphisms from repetitive regions also impact the observed demographic history?

**Description of the work carried out during the visit**

The dataset consisted of whole-genome shotgun sequences from three species of Old World camel: *C. dromedaries* (n = 9), *C. ferus* (n = 9), and *C. bactrianus* (n = 7). Paired-end sequencing of each individual was performed on a single lane of an Illumina HiSeq with a mean insert size of 500 bp. Reads were 3’ trimmed to a minimum phred base quality score of 20 and minimum length of 50 bp using POPOOLATION v1.2.2 (Kofler et al. 2011). Trimmed reads were mapped to the *C. ferus* CB1 genome assembly (Genbank ID GCA\_000311805.2) using BWA v0.6.2 (Li and Durbin 2009). For *C. dromedarius*, reads were also mapped to an in-house *de novo* assembly of the dromedary genome. All alignments were filtered to contain only high quality (MQ ≥ 20), unique (MarkDuplicates; PICARD v1.89, http://picard.sourceforge.net), unambiguously mapped, and properly paired reads (SAMTOOLS v0.1.19, Li et al. 2009). All single nucleotide polymorphisms (SNPs) and corresponding genotypes used in downstream analyses were identified using the GATK Best Practices Pipeline (Van der Auwera et al. 2013).

To model the demographic history, we first used the Pairwise Sequentially Markovian Coalescent model (PSMC; Li and Durbin 2011). Analyses were performed according to default parameters with a minimum and maximum coverage cutoff per base set to 5 and 40, respectively, and with 100 bootstrap replicates. Results were plotted assuming a generation time of 5 years for camels. To remove the potential effects of repetitive elements on the distribution of heterozygous sites, we constructed a consensus genome sequence for each individual using the ‘FastaAlternateReferenceMaker’ walker in GATK using our cleaned genotype data (VCF file). Subsequently, the repetitive regions in the resulting genome were masked using the annotations available for the CB1 assembly. The PSMC analysis was then repeated as described above.

Because the method implemented in PSMC lacks power to detect recent (less than ~2000 years ago in camels) changes in effective population size (Ne) due to the lack of recombination events (Li and Durbin 2011), we used the software SNeP (Barbato et al., under development) to infer more recent demographic history. SNeP uses the extent of linkage disequilibrium (LD) between markers at various distances to infer Ne­­. For each species of camel we included SNPs found in the 5000 largest scaffolds that were polymorphic (minimum allele frequency ≥ 0.05), had a genotyping rate ≥ 0.9, and did not deviate from Hardy-Weinberg equilibrium (HWE; *p* ≥ 0.0001) using PLINK v1.9 (https://www.cog-genomics.org/plink2). Because the computational time necessary for SNeP increases exponentially with the number of SNPs, we determined the threshold where increasing the number of SNPs did not produce significant decreases in variance of Ne. We generated 10 random subsets of 5,000, 10,000, 20,000, 50,000, and 100,000 SNPs in *C. ferus*, phased the genotype data using BEAGLEx v3.3.2 (Browning and Browning 2007) and ran SNeP using the method of Corbin et al. (2010). The previous pipeline was replicated 10 times in each species for the threshold number of SNPs identified.

**Description of the main results obtained**

The mean coverage of each genome alignment was approximately ~15X. In total, across all three species of camels 4,960,087 SNPs were called. Insertion/deletion polymorphisms were excluded from downstream analyses. The results of the demographic history recreated from raw genome alignments using PSMC can be found in Figure 1. All three species appeared to have suffered an extended period of relatively small Ne between 200k – 20k years ago. This population reduction may be a result of the last glacial maximum associated with the late Pleistocene between 120k – 12k years ago. A similar pattern has been observed in other mammals from the Northern Hemisphere (e.g. Zhao et al. 2012). In the wild camel, *C. ferus*, a rapid population expansion occurred potentially following the last glacial maximum but then abruptly ended and a second bottleneck followed, probably eventually resulting in the small Ne observed today. The domestic *C. bactrianus* shared a similar ancestry with *C. ferus* until 20k years ago, where it experienced a less substantial expansion and eventual reduction. The most recent bottleneck, however, appeared to have been earlier and less severe than in *C. ferus*. Although speculative at this point, contact with early humans in the region may have initiated this decline prior to domestication ~6k years ago. Finally, in dromedaries, a similar pattern also emerged, albeit the events occurred more recently and N­e was consistently smaller than in the other two species. It is possible that parameters such as generation time and mutation rate are quite different in *C. dromedarius* compared with the other species, and these factors need to be investigated further.

We also investigated the potential effects of two technical concerns: 1) mapping to a relatively divergent reference genome and 2) the effect of removing SNPs from repetitive regions. In the former, we mapped our *C. dromedarius* reads to both the *C. ferus* reference genome and a *de novo* dromedary genome in preparation in our laboratory (Figure 1 dark blue and light blue lines, respectively). Despite a divergence time of ~5 million years ago between *C. dromedarius* and *C. ferus*, the effects on historical Ne were rather minimal. Mapping to the more divergent genome appears to consistently underestimate Ne, but the exact mechanism for this remains to be determined. Several factors such as genome assembly quality, repeat structure, SNP-calling algorithms, etc. may influence this result. In the latter, the removal of repetitive regions prior to analysis with PSMC results in similar patterns but with several important differences (Figure 2). First, the absolute estimates of effective population size are much lower than the previous estimates and uninformative prior to ~13k years ago; probably a result of the fewer number of heterozygous bins thus fewer coalescent events. Second, the most recent population decline in both *C. ferus* and *C. bactrianus* occurred several thousand years more recently than in Figure 1, congruent with the current archaeological estimates of initial domestication.

Because PSMC lacks resolution in the recent (last few thousand years) past, we used the program SNeP to explore patterns in Ne using LD. Using just *C. bactrianus*, we found that 20k SNPs was sufficient to estimate Ne and increasing the number of SNPs did not gain any resolution in demographic history but rather resulted in large increases in computational time (Figure 3). Using permutations of 20k SNPs, we observed a similar population trajectory indicating a gradual decline in all three camel species over the last 1,000 generations (~5,000 years) (Figure 4). This is in agreement with the estimates of PSMC and expectations of domestication. However, using SNeP, *C. ferus* consistently had a smaller Ne than the other species, which was not observed in the PSMC results. This methods, however, assumes accurate estimates of allele frequencies to infer LD within a population, and our dataset, which is based on a small sample size (n = 7 or 9), may be too small to reliably estimate LD. Furthermore, it is possible our dataset contains individuals from structured populations, which may also confound calculations of LD.

**Future collaboration with the host institution**

The interaction and collaboration with the host was a fruitful experience. We plan on staying on close contact because both groups are working on similar questions and analyses in two different groups of camelids. Additionally, we will remain in close contact in order to further test the utility of the software SNeP, which is being developed in the Bruford Lab. Finally, the collaboration will continue in a larger project investigating the patterns of demographic history in vertebrates in general from whole genome sequencing.

**Projected publications / articles resulting or to result from the grant** (*ESF must be acknowledged in publications resulting from the grantees work in relation with the grant*)

Our future plan is for these results to incorporated into two future publications, one focused on our new dromedary genome assembly and a second on demographic history and selection in the other two species (*C. ferus* and *C. bactrianus*). In both cases the ESF and the host institution will receive proper acknowledgement for their support.

**Other comments (if any)**

I am extremely grateful to Dr. Mike Bruford and the Cardiff University School of Biosciences not only for their support during my stay, but providing important information necessary when encountering visa complications. I would also like to thank all members of the Bruford Laboratory, especially Dr. Pablo Orozco-terWengel and Mario Barbato for the help I received from them. Finally, thanks to Jukka Corander and the CSC — IT Center for Science Ltd in Finland for use of their computational resources.



**Figure 1:** Demographic history of three camel species up to 1,000 years ago reconstructed using PSMC (Li and Durbin 2011). Each line represents the estimated effective population size (N­e) for an individual genome according to the colors depicted in the legend. Bootstrap replicates are not shown for clarity, but variance is quite large more recently than 103 and older than ~5x105 years ago.

Abbreviations: g = generation time, u = mutation rate.



**Figure 2:** Historical effective population sizes of each camel species recreated using PSMC (Li and Durbin 2011) and a set of genome-wide SNPs filtered to exclude annotated repetitive regions. Each line represents an individual genome according to the colors depicted in the legend. Bootstrap replicates are not shown for clarity, but variance is quite large more recently than 103 and older than ~5x105 years ago.

Abbreviations: g = generation time, u = mutation rate.



**Figure 3:**  Estimates of Ne from SNeP. Each line represents a different permutation of the number of SNPs indicated by the color in the legend. There is little improvement when using 20k SNPs or more.



**Figure 4:**  Estimates of Ne from SNeP for each species of camel (see legend for colors). Each line represents a different permutation of 20k SNPs.

**References**

Browning, S.R. and Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* 81, 1084-1097.

Corbin, L.J., Blott, S.C., Swinburne, J.E., Vaudin, M., Bishop, S.C. and Woolliams, J.A., 2010. Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Animal Genetics* 41, 8-15.

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

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

Li, H. and Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475, 493-496.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.

Rosenberg, N.A. and Nordborg, M., 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics* 3, 380-390.

Van der Auwera, G.A., Carneiro, M.O., et al. 2002. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 11:11.10:11.10.1–11.10.33.

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., Hu, Y., He, W., Zhang, S., Fan, W., Zhu, L., Li, D., Zhang, X., Chen, Q., Zhang, H., Zhang, Z., Jin, X., Zhang, J., Yang, H., Wang, J., Wang, J. and Wei, F., 2012. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nature Genetics* 45, 67-71.