Research Networking Programmes

Short Visit Grant ☒ or Exchange Visit Grant ☐

(please tick the relevant box)

Scientific Report

The scientific report (WORD or PDF file – maximum of eight A4 pages) should be submitted online within one month of the event. It will be published on the ESF website.

***Proposal Title*: Paleogenomics of Spitsbergen bowhead whales (*Balaena mysticetus*)**

Magdalena Gonciarz, Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Poland

Final Scientific Report, ESF Research Networking Programmes, Short visit at Natural History Museum, Oslo, Norway under supervision of Professor Lutz Bachmann

Application Reference N°: 6333

1) Purpose of the visit

Objectives

The aims of my stay at NHM, Oslo were following:

1. Learning bioinformatics methods for the assembly of NGS data with particular emphasis on obtaining whole mitochondrial genomes (MITObim software, developed at NHM Oslo for such purpose)
2. Analyzing the available NGS data for addressing the genetic structure and temporal changes of the historic Spitsbergen stock of bowhead whales (parameters that are of particular importance for conservation
3. Familiarizing with the special challenges of working with ancient, degraded DNA samples and quality assessment of such data

2) Description of the work carried out during the visit

Introduction

Bowhead whale (*Balaena mysticetus*) is large (approximately 20 m) baleen whale (*Mystycete*) that inhabits arctic and subarctic regions **(IUCN, Fig. 1)**. It is considered as a relict species both geographically and evolutionary **(Finley 2001)**. Bowhead whale is a long – living species (up to 200 years). Being the filter feeder, its diet consists of high amounts of copepods (mainly *Calanus*) **(Finley 2001)**.

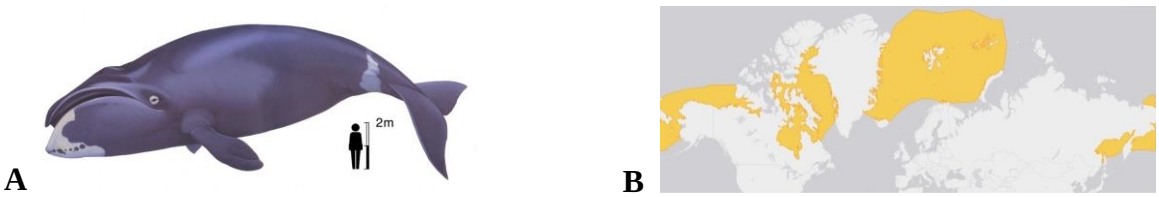


Fig.1. Bowhead whale (*Balaena mysticetus*) (A) and its geographical rage (B)

It is believed that this species has arisen in the Northern Hemisphere in the Pleistocene (5.5 – 1.6 million years ago) **(Finley 2001)**. During the last ice age, bowhead whales occurred in the Gulf of St. Laurence (Canada), and during the climatic optimum (10,000 – 7500 years before present, BP) they spread over the Canadian Arctic Archipelago, possibly seasonally colonizing the eastern Beaufort and Bering Seas **(Finley 2001, Rekdal et al. 2015**). The Spitsbergen stock (Svalbard, Norway) presumably originates from a refugium in the Eastern North Atlantic **(Wiig et al. 2010)**.

Bowhead whales exhibit strong site fidelity. Therefore large proportion of the population can be observed in few locations where it forms congregations or some of the individuals may pass by **(Finley 2001, Wiig et al. 2011)**. Moreover, populations show seasonal distribution. During winter individuals migrate south possibly via one of known migration corridors, which is Baffin Island. The spring migration path would be Lancaster Sound when bowheads migrate north while temperature rises **(Finley 2001)**.

The peak in abundance of these marine mammals was reached 10,000 and 8500 BP. However, early in 1600s in the North Atlantic region near Svalbard, intensive commercial whaling of bowheads, was initiated **(Finley 2001, Wiig et al. 2011)**. Spitsbergen stock was one of the most exploited and today it is considered Critically Endangered **(IUCN)**. Before this over exploitation stock size was estimated at 25,000 up to over 100,000 individuals. But at the moment, sampling can hardly be performed **(Wiig et al. 2010)**.

Due to antropopression, International Whaling Commission (IWC 1978) has stated that it is crucial and urgent to reduce human – induced mortality of these animals **(Finley 2001, Wiig et al. 2011)**. Conservation concerns resulted in extensive analyses of the population size and past history of bowhead whales population using various approaches **(Finley 2001, Sasaki et al. 2005, Heide – Jorgensen et al. 2010, Wiig et al. 2010, Wiig et al. 2011, Lyndersen et al. 2012)**. A collection of well preserved ancient bowhead whale bones from Svalbard is deposited at the Natural History Museum (NHM) in Oslo, Norway. Due to its relatively high copy number and higher likelihood for DNA survival, mitochondrial DNA (mtDNA) is one of the preferred genetic markers applied not only in contemporary samples but also in ancient DNA analyses. In population genetic studies and phylogeography, the control region (CR), exhibiting high levels of inter - population variability, has been commonly used **(Borge et al. 2007, Hahn et al. 2013)**.

Based on sequencing of CR region from historical samples from depleted Spitsbergen and the Bering/ Chukchi/ Beaufort (BCB) bowhead whale stocks, authors suggested lack of significant genetic differentiation between two of them **(Borge et al. 2007)**. However, they argue that using other molecular markers (e.g. nucleotide sequences from more conserved regions of mtDNA), could provide more information on bowhead whale stock characteristics. On the other hand, the reconstruction of the

phylogeny of the baleen whales was proposed by **Sasaki et al. 2005**. Authors determined the complete mitochondrial genome sequence for 10 contemporary living baleen species (including bowhead whale), using direct sequencing and primer walking methodology. But so far, the study on genetic structure and demographic history of pre – exploit Spitsbergen stock based on whole mitogenome analyses have not been proposed. Such data could provide new insights in the future conservation and management practices of these animals.

Recently, the Next - Generation Sequencing (NGS) became a straightforward alternative for laborious analyses of multilocus markers for non – model organisms in phylogeography and population genomics. Application of NGS could be more time – efficient and also cost – effective **(McCormack et al. 2013)**. Moreover NGS has been shown to provide data that can substantially increase resolution and provide better topologies and divergent time estimates than shorter mtDNA sequences such as the CR. The NGS approach may be applied in determination of deep and more recent radiations, but could be also very efficient in understanding inter – and intra – specific diversity or evolutionary patterns. It is worth noticing, that NGS technique may be very useful in ancient DNA (aDNA) studies since it perfectly to the physicochemical properties of ancient DNA molecules. Again, this is very useful for better resolution of phylogeny and phylogeography **(Hancock et al. 2013 and references therein)**. One of the well - known problems, while working with aDNA is its poor quality, manifested *inter alia* by the fragmentation of the DNA. However it is not an issue in Next Generation Sequencing, because the DNA template in NGS has to be first cut into shorter DNA fragments. Despite that features, NGS has a quite good throughput and consequently is less expensive for large - scale sequencing. Therefore could be well applied in paleogenomics **(Rizzi et al. 2012).**

Although, the analysis of NGS outputs may seem very challenging and resource demanding task, it currently becomes more available thanks to computer tools helping overcome the huge data sets. There are few software's specializing in whole genome assemblies. MITObim **(Hahn et al. 2013)** program is one of the examples, recently becoming more popular, especially thanks to its specificity. It can be applied for mixed species samples and has provided easy to use script wrapper. Thus, it is an easy to use alternative for those researchers with limited skills in bioinformatics **(Hahn et al. 2013)**.

Materials & Methods

The ^{14}C dated samples of old bowhead whale bones were deposited at NHM, Oslo. The DNA extraction was performed earlier and excellent DNA survival was shown. Next - Generation Sequencing (NGS) data were already obtained for 14 historical bowhead whale samples.

Program MITObim **(Hahn et al. 2013)** was used for whole mitochondrial genome assembly of 14 bowhead whale samples. The NGS raw data sequences sets (reads) obtained for 10 samples were already processed for direct downstream analyses. For the other samples, I repeated the procedure of concatenation and interleaving using Unix® system. All 14 interleaved output sequences served then as templates for MITObim analyses. One assembly strategy was using complete mitochondrial genome sequence of bowhead whale (GenBank Accession number: AP006472.1, **http://www.ncbi.nlm.nih.gov**) as a reference. Additionally for one of the samples the significantly shorter sequence of COI (GeneBank Accession number: AP006472.1) was

used as a seed for the assembly. The aim of this analysis was a comparison of both assembly results (from whole mirochondrial genome reference and shorter sequence reference approaches).

In order to view the assembly results and evaluate the coverage depth of the assembled sequence the software Tablet 1.14.10.20 **(Milne et al. 2013)** was used. It was also possible to preliminary evaluate the probability of contamination of the samples with other then bowhead whale DNA material. It was done by comparing the samples, which presented very poor mitochondrial DNA coverage. The assumption was that if exactly the same coverage pattern would occur in compared samples, it might indicate contamination with human DNA. Additionally, new assembly was performed for randomly selected samples with human mitochondrial genome (GenBank Accession number: AP008824.1) as a reference.

The assembled contigs were subjected to Blast **(http://blast.ncbi.nlm.nih.gov)** searches against GenBank data base to confirm identity. The MITOS **(Bernt et al. 2013)** and DOGMA **(Wyman et al. 2004)** web servers were used for automated annotation of the obtained mitogenomes.

Using various phylogenetic computer tools: Clustal X 2.0 **(Larkin et al. 2007)**, BioEdit **(Hall 1999)** and MEGA 6 **(Tamura et al. 2013)** 14 assembled genomes were aligned aiming to observe genetic diversity of studied samples.

3) Description of the main results obtained

Results & Discussion

14 mitochondrial genomes, representing historical samples from the Spitsbergen stock of bowhead whales were successfully assembled using MITObim software with a whole mitochondrial genome sequence of bowhead whale retrieved from GenBank as a reference. Doing so, I have learned handling different analyses programs needed for whole genome assembly and became familiar with their specific features. I was also able to preliminary evaluate the quality of obtained data, performing few additional analyses.

MITObim software uses a baiting and iterative mapping approach and employs the respective MIRA v3.4.1.1 modules **(Chevreux et al. 1999)**. It is especially focused on the assembly of whole mitochondrial genomes of non – model organisms with only distantly related reference sequences at hand. However it also gives the possibility to use either short reference sequence or complete mitochondrial genome as a bait (*seed*), which makes the analysis usually less time consuming and more straightforward. I performed both such analyses to confirm whether the resulting mitochondrial sequences would be identical:

1. Using whole mitochondrial genome of bowhead whale as a reference, the average number of iterations equaled two and the MITobim completed assembly with resulting sequence of 16,776 bp
2. Using COI gene sequence of bowhead whale as a reference, the four separate analyses had to be performed. The procedure was as follows:
   - In each of programs runs, the assembling process has been stopped, once the elongated sequence reached its maximum possible length
   - This novel assembled incomplete sequence was a template for the next run, and the process would be repeated, until the MITObim completes the assembly

- The resulted sequence maximum length was 11,946 bp and was incomplete in comparison to the reference bowhead mitochondrial genome (16,389 bp). In order to try to feel in the gap, the assembly was repeated using the fragment of the reference sequence, which was missing

The fragment between 1 – 4689 nucleotide of reference sequence could not be finally assembled at this stage of analysis. The annotation of the complete mitochondrial genome of bowhead whale indicates that the assembly process stops on the 757 position of the ND2 gene and it should be further reconsidered.

The obtained coverage of 14 complete mitochondrial sequences was satisfactory for most of the studied samples, on average was equal to 25,2. Four samples exhibited very low coverage and probably would be excluded from analysis based on whole mitogenome approach. The additional NGS data were earlier obtained (re - run) and the analysis I performed indicates the increase of coverage depth of this samples **(Tab.1)**. Moreover, two - three iterations would be enough to complete the assembly when using whole mitochondrial genome as a reference. Such approach was very time efficient and straightforward. The average length of the complete mitochondrial genome of bowhead whale obtained was 16,7484 bp **(Tab.1)**.

Tab.1. The sample name, reference sequence used, final length, average coverage in bp and number of iterations of 10 bowhead whale NGS data sets

Sample Name	Reference sequence used	Final contig lenght	Coverage	Number of iterations
315 a	complete mtDNA	16,681	9,7	3
315 b	*complete mtDNA*	*16,671*	*32,1*	*2*
316 a	complete mtDNA	16,762	15,3	3
316 b	*complete mtDNA*	*16,766*	*43,1*	*2*
317	complete mtDNA	16,774	14,7	2
318	complete mtDNA	16,679	2,3	2
319	complete mtDNA	16,734	2,7	2
320	complete mtDNA	16,756	7,9	4
321	complete mtDNA	16,778	37,6	3
322	complete mtDNA; COI	16,776	128,3	2*
323	complete mtDNA	16,778	50,5	3
324	complete mtDNA	16,766	14,3	3
152	complete mtDNA	16,739	13,3	2
153	complete mtDNA	16,679	9,5	2
167	complete mtDNA	16,78	10,1	4
169	complete mtDNA	16,756	12,3	3

* complete mtDNA analysis
complete mtDNA - GenBank Accession number: AP006472.1
COI - GeneBank Accession number: AP006472.1
a – the first run; b – the re – run combined with the first run (italics)

There is a variety of steps, which may help to estimate the quality of the data obtained from NGS. One of the threats, while working with aDNA is contamination of the samples, with other DNA material. The most often observed one would be human

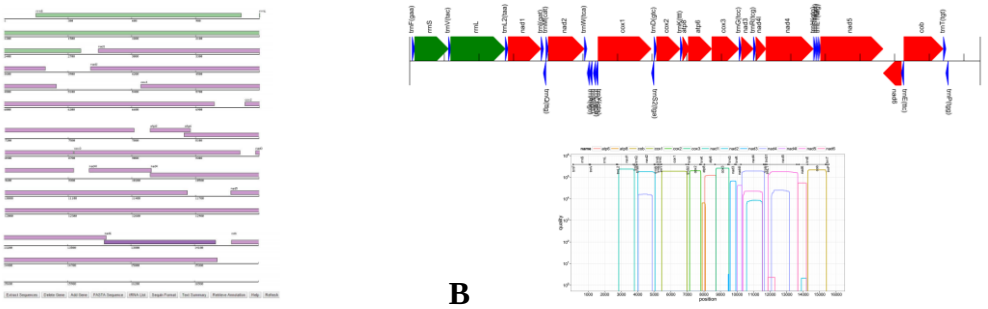
genetic material, originating from the examiner who extracted DNA, but also from others who had previously contact with the sample. Additionally to laboratory controls carried out previously, two preliminary evaluations which indicate lack of contamination were performed. First, the assembly in MITObim was done based on the human mitochondrial DNA sequence as a bait. The obtained result supports the assumption of lack of contamination. If human DNA material would be detected in the sample, then one or two iterations would be enough to complete the assembly of this mitochondrial genome and the result could be easily identified via blast search. None of such effects have been observed for studied bowhead sample. Secondly, the other test was performed, based on the comparison of the samples with poor coverage. It is in accordance with latter control. Therefore there is low probability of contamination in the studied samples. The comparison of high and low coverage is shown in **Fig.2**.

There are other analyses, which may be helpful in evaluation of the data quality. One of them would be to do the blast searches using all NGS reads, to estimate the quantity of reads which would match other than target DNA (depending on organism studies e.g. fish, fungi, bacteria etc.). There are also quality scores provided for NGS data and the method on its own
is considered quite robust. One would expect some problematic regions, especially in the end of each read. Therefore the trimming would be recommended generally.

A **B**

Fig. 2. Example of high (A) and low (B) coverage depth of Next – Generation Sequencing data visualized using Tablet 1.14.10.20 software for two historical samples of bowhead whale

The core annotation of mitochondrial genome of bowhead whale obtained with NGS method was performed. Two different annotation methods applied, show the same order of the protein genes **(Fig. 3)**, also with accordance to previous results obtained with primer walking sequencing **(Sasaki et al. 2005)**.



A **B**

Fig.3. Preliminary core annotation of mitochondrial bowhead whale genome obtained using Next – Generation Sequencing, using DOGMA (A) and MITOS (B) web servers

4) Future collaboration with host institution (if applicable)

The future plans would consider further genetic analyses such as: estimation of the mutations rates for different regions of mitochondrial genome (evolutionary rates of genes), looking for genetic adaptation and reconstructing phylogeny (phylogenetic trees). Moreover, it would be possible to identify different haplotypes, estimate their frequency and relation (e.g. drawing a network). Having such information, the genetic differentiation between contemporary and ancient stocks of bowhead whales could be measured. The reconstruction of the past demography events (especially bottlenecks) and also patterns of isolation of different stocks would be one of the most interesting objectives.

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

The obtained results may provide a part of the materials, which could be subsequently included in the scientific paper.

6) Other comments (if any)

Conclusions

During my stay in NHM, Oslo I realized following objectives:

1. Having Next – Generation Sequencing data and using MITObim software with Unix® system, I successfully assembled 14 mitochondrial genomes of ancient bowhead whales
2. In order to evaluate the quality of obtained results, I used several different approaches of data analysis with special focus on aDNA
3. To conclude and present suggested plans for the future work, I reviewed the most recent literature regarding analyzed data

Acknowledgements

I express my gratitude to Professor Lutz Bachman for help, encouragement and all his suggestions during my stay at NHM, Oslo. Opportunity for completing this visit granted by European Science Foundation is gratefully acknowledged.

References

1. Bernt, A. D., Jühling F., Externbrink F., Florentz C., Fritzsch G., Pütz J., Middendorf M., Stadler P. F., 2013, MITOS: improved de novo Metazoan mitochondrial genome annotation, Molecular Phylogenetics and Evolution, Vol. 69, No. 2, 313 – 319
2. Borge T., Bachmann L., Bjornstad G., Wiig O., 2007, Genetic variation in Holoce bowhead whales from Svalbard, Molecular Ecology, Vol. 16, 2223 - 2235
3. Chevreux B., Wetter T., Suhai S., 1999, Genome sequence assembly using trace signals and additional sequence information, Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)
4. Finey K. J., 2001, Natural history and conservation of the Greenland Whale, or Bowhead, in the Northwest Atlantic, Arctic, Vol. 54, No 1, 55 – 76
5. Hahn C., Bachmann L., Chevreux B., 2013, Reconstructing mitochondrial genomes directly from genomic next – generation sequencing reads – a baiting and iterative mapping approach, Nucleic Acids Research, 1 – 9

6. Hall T. A., 1999, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, Nucleic Acids Symposium Series, Vol. 41, 95 - 98
7. Hancock – Hanser B. L., Frey A., Leslie M. S., Dutton P. H., Archer F. I., Morin P. A., 2013, Targeted multiplex Next Generation Sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics, Molecular Ecology, Vol. 13, 254 – 268
8. Heide – Jorgensen M. P., Laidre K. L., Wiig O., Postma L., Dueck L., Bachmann L., 2010, Large – scale sexual segregation of bowhead whales, Endangered Species Research, Vol. 13, 73 – 78
9. IUCN, 2015, http://www.iucnredlist.org/details/2467/0
10. Larkin M. A., Blackshields G., Brown N. P., Chenna R., McGettigan P. A., McWilliam H., Valentin F., Wallace I. M., Wilm A., Lopez R., Thompson J. D., Gibson T. J., Higgins D. G., 2007, Clustal W and Clustal X version 2.0. Bioinformatics, Vol. 23, 2947 – 2948
11. Lyndersen C., Freitas C., Wiig O., Bachmann L., Heide – Jorgensen M. P., Swift R., Kovacs K. M., 2012, Lost highway not forgotten: satellite tracking of a bowhead whale (Balaena mysticetus) from critically endangered Spitsbergen stock, Arctic, Vol. 65, No., 1, 76 – 86
12. McCormack J. E., Hird S. M., Zellmer A. J., Carstens B. C., Brumfield R. T., 2013, Applications of Next Generation Sequencing to phylogeography and phylogenetics, Molecular Phylogenetics and Evolution, Vol. 66, 526 – 538
13. Milne I. S. G., Bayer M., Cock P. J. A., Pritchard L., Cardle L., Shaw P. D. and Marshall D., 2013, Using Tablet for visual exploration of second - generation sequencing data, Briefings in Bioinformatics Vol. 14, No. 2, 193 – 202
14. Rekdal S. L., Hansen R. G., Borchers D., Bachmann L., Laidre K. L., Wiig O., Nielsen N. H., Fossette S., Tervo O., Heide – Jorgensen M. P., 2015, Trends in bowhead whales in West Greenland: aerial surveys vs. genetic capture – recapture analyses, Marine Mammal Science, Vol. 31, No. 1, 133 – 154
15. Rizzi E., Lari M., Gigli E., De Bellis G., Caramelli D., 2012, Ancient DNA studies: new perspective on old samples, Genetics Selection Evolution, 44 – 21
16. Sasaki T., Nikaido M., Hamilton H., Goto M., Kato H., Kanda N., Pastene N. A., Cao Y., Fordyce R. E., Hasegawa M., Okada N., 2005, Mitochondrial phylogenetics and evolution of *Mysticete* whales, Systematic Biology, Vol. 54, No. 1, 77 – 90
17. Tamura K., Stecher G., Peterson D., Filipski A., Kumar S., 2013, MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0., Molecular Biology and Evolution, Vol. 30, 2725 – 2729
18. Wiig O., Heide – Jorgensen M. P., Lindqvist C., Laidre K. L., Postma L. D., Dueck L., Palsboll P. J., Bachmann L., 2011, Recaptures of genotyped bowhead whales Balaena mysticetus in eastern Canada and West Greenland, Endangered Species Research, Vol. 14, 235 – 242
19. Wiig O., Bachmann L., Oien N., Kovacs K. M., Lydersen C., 2010, Observations of bowhead whales (Balaena mysticetus) in the Svalbard area 1940 – 2009, Polar Biology, Vol. 33, 979 – 984
20. Wyman S. K., Jansen R. K., Boore J. L., 2004, Automatic annotation of organellar genomes with DOGMA, Bioinformatics, Vol. 20, No. 17, 3252 – 3255