European Molecular Biology Laboratory
Laboratoire Européen de Biologie Moléculaire
Europäisches Laboratorium für Molekularbiologie

Dr Vladimir Beneš

EMBL Genomics Core Facilities

Charles University in Prague
Faculty of Science
Department of Student Affairs
Albertov 6
CZ-128 43 Prague 2, the Czech Republic

Heidelberg, 19 June 2020

Subject: Review of the Ph. D. thesis "Analysis of the genome of a free-living amoeba *Mastigamoeba balamuthi* and its comparison with pathogenic *Entamoeba histolytica*" submitted for its defense by Mr. Vojtěch Žárský, M. Sc.

To whom it may concern

The doctoral thesis submitted by Vojtěch Žárský reports on (i) analyses of the genome of *M. balamuthi* and its comparison with Entamoebidae genomes, (ii) elucidation of processes leading to adaptation of *M. balamuthi* to a free-living anaerobe, (iii) identification of genomic features enabling evolution of Entamoebidae into a parasitic group, and (iv) investigation of the role and distribution of peroxisomes in Archamoebae and other eukaryotes.

Vojtěch Žárský's relevant and ultimately successful work, conducted under supervision of Professor Jan Tachezy, focuses on sequencing and analyses of the *M. balamuthi* genome and transcriptome. It stands in the centre of his PhD thesis and also represents a starting point for other analyses included in his dissertation.

The *M. balamuthi* genome was sequenced using 454 Roche sequencing and short-read Illumina sequencing technologies. Employed Bioinformatics algorithms are appropriate with the MaSuRCA toolkit for DNA genome assembly and Augustus, Blast2GO and InterProScan for gene prediction and annotation, aided by the additional mRNA-Seq data generated for the *M. balamuthi* transcriptome. BUSCO was used to assess the completeness of the assembly and annotation process.

Genome assembly projects are among the most difficult Bioinformatics analysis tasks and the presented results suggest that Vojtěch Žárský has managed to accurately assemble an advanced draft genome of *M. balamuthi*. Assembly parameters (1925 scaffolds (N50=442.5 kbp), an estimated genome size of 57.27 Mbp, 82.8% of conserved genes completely assembled (BUSCO validation) and 98.6% 'mappable' RNA-Seq reads using GMAP) are impressive and although still a draft it definitely provides a valuable resource not only for parasitologists. The genome has been duly deposited in specialized evolutionary genomics database (https://bioinformatics.psb.ugent.be/orcae/overview/Masba) and the European Nucleotide Archive to serve the wider scientific community. I would have liked to see a more thorough evaluation of alternative assembly algorithms (e.g., SPAdes, Newbler) and full-fledged assembly annotation pipelines (e.g., MAKE) but these are minor deficiencies.

Presented evolutionary analysis made use of orthologous groups in EggNOG and the hidden Markov model (HMMER) algorithms. Multiple sequence alignments were constructed using MAFFT and the BLOSUM30 scoring matrix for distantly related amino acid sequences. These choices of algorithms and database are justifiable and appropriate. Vojtěch Žárský additionally explored and compared mobile elements (DNA transposons and retrotransposons) in *M. balamuthi* and two *Entamoeba* species, however with quite different results. It remains unclear if this is true

Meyerhofstrasse 1
D-69117 Heidelberg
Germany

Phone: +49-6221-387-8486
FAX: +49-6221-387-8306
E-mail: benes@embl.de
URL: www.genecore.embl.de

European Molecular Biology Laboratory
Laboratoire Européen de Biologie Moléculaire
Europäisches Laboratorium für Molekularbiologie

Dr Vladimír Beneš

EMBL Genomics Core Facilities

biology or related to different assembly approaches. The hybrid assembly approach taken for *M. balamuthi* with 454 and Illumina data is likely superior for long LINE1 elements compared to assemblies from short-read data alone. Vojtěch Žárský's effort in this regard is crowned by the first-author manuscript, which has been submitted to the 1st tier journal Molecular Biology and Evolution (impact factor for 2019: 14,7).

Another main project presented in this thesis is the comparison of predicted proteomes of 111 metazoans with respect to peroxisomes. Vojtěch Žárský analysed the presence or absence of 14 peroxins conserved in metazoans using hidden Markov models. Interestingly, the identified metazoan lineages lacking peroxisomes have very different phylogenetic positions and the loss of peroxisomes seems to be associated with a general pattern of genome shrinkage and reductive evolution. Peroxin homologs were carefully checked in all 6 frames and in the transcriptomic data which appears technically sound. Notably, this analysis is entirely based on *in silico* predicted proteomes from genomic data. Presented argument could have been strengthened by direct proteomic read-outs.

The thesis itself is lucidly written but some Figures could have been improved; also their reproductions from the literature could have been avoided.

All in all, Vojtěch Žárský has generated interesting research results with the appropriate scientific rigor and accuracy. Given the complexity of the various Bioinformatics projects he has worked on (*i.e.*, *de novo* genome assembly, genome annotation, phylogeny and evolutionary genomics) I do consider his thesis very good. My above points do not terribly reduce its quality.

In summary, Vojtěch Žárský has successfully pursued the challenging and interesting PhD project and achieved its goals. In my view, he has convincingly demonstrated a scientific maturity expected from a successful PhD candidate and I recommend awarding the PhD title to him.

With best regards,

Dr Vladimír Beneš
Head of EMBL Genomics Core Facilities

Meyerhofstrasse 1
D-69117 Heidelberg
Germany

Phone: +49-6221-387-8486
FAX: +49-6221-387-8306
E-mail: benes@embl.de
URL: www.genecore.embl.de