Peer-Review Record:

A Manual Curation Strategy to Improve Genome Annotation: Application to a Set of Haloarchael Genomes

Friedhelm Pfeiffer * and Dieter Oesterhelt

Reviewer 1: Gerald Schönknecht
Reviewer 2: Anonymous
Editors: Hans-Peter Klenk, Michael W. W. Adams and Roger A. Garrett (Guest Editor of Special Issue "Archaea: Evolution, Physiology, and Molecular Biology")

First Round of Evaluation

Round 1: Reviewer 1 Report and Author Response

The manuscript by Pfeiffer and Oesterhelt describes the problem of "over-annotation" of genome/protein databases and its propagation by automated annotation pipelines. The authors describe an alternative, manual curation strategy aimed at avoiding over-annotation.

While not especially flashy or exciting, the authors tackle a serious problem, the reliability of annotations in publically available databases. The authors are to be complimented for making the effort to provide well curated annotations for a set of haloarchaeal genomes, and contribute to improving public databases (EMBL/Genbank, UniProt). The manuscript describes in detail the approach used by the HaloLex genome annotation system to obtain reliable, consistent annotations, based on "Gold Standard Proteins". The manuscript is well written and good to read, and the only point of critique this reviewer would like to raise is that some of the examples given are not easy to follow.

Giving specific examples is certainly a good idea. However, being experts in their field the authors might not be aware that for readers less familiar with the metabolism/physiology of archaea, the examples are not always easy to follow, such as examples given on p. 7 starting line 196 (The main difference here is between a **ferredoxin**-dependent and/or a **NAD-**dependent decarboxylation of pyruvate into acetyl-CoA?/A few lines later when I read that "halophilic archaea do not contain the coenzyme methanopterin", I had to go back and reread the preceding sentences to get an idea of the argument.), p. 10 starting Line 299, and p. 12 starting Line 354.

Response: We have rephrased the four paragraphs where the referee found that the described examples are not always easy to follow.

Round 1: Reviewer 2 Report and Author Response

I received this manuscript and looked at it with joy. So often people try to publish quickly "dirty data", and later someone needs to polish things up. This is a great paper to aim at such a direction. All efforts to clarify and systemize data are welcome, and therefore this paper is very important and should be published. The curation procedure is done, to my mind, very well; and I am confident that this will help the archaea community enormously.

Importance: Archaea is/are the third domain of life, no more or less!

I did not find problems in writing (narrative very fluent) or in the logic. As two of the top researchers in the archaea community had read the manuscript and given critical comments, I am also very confident that the details of the genes discussed are adequate.

There is now available a great resource for the gene discovery in Haloferax volcanii (Kiljunen et al 2014: BMC Biology). A comprehensive insertion mutant library has been generated for this model archaeon in order to facilitate gene discovery. This should be mentitioned and discussed in the text; *i.e.*, how this resource could be used to assign functions for genes.

In summary, this paper is of good quality and of enormous value, and should be published.

Response: We have added a short "outlook-type" paragraph towards the end of the conclusions. In this paragraph we mention three approaches that will facilitate gene discovery, among them large-scale screening techniques as exemplified by the insertion mutant library pointed out by the referee.