Management Plan for the Annotation of Cryptosporidium baileyi

A Data Management Plan created using DMPTool

Creator: Shelton Griffith

Affiliation: University of Georgia (UGA)

Template: National Institutes of Health (NIH)

Last modified: 12-11-2015

Management Plan for the Annotation of Cryptosporidium baileyi

Data sharing plan

Data Management Plan

Products of Research

Globally, one in ten child deaths result from diarrheal disease during the first 5 years of life, resulting in nearly a million fatalities worldwide annually. One of the main culprits of this horrible statistic is the zoonotic apicomplexan protist, *Cryptosporidium*. Parasites of the genus *Cryptosporidium* infect a wide range of vertebrates, from fish to humans and some species are capable of zoonotic transmission. Before the beginning of the AIDS epidemics, these parasites were not considered important pathogens however that narrative has quickly changed and they are now recognized as among the most common enteric infections in humans and livestock. In immune-competent individuals the infection is characterized by mild to severe self-limiting diarrhea but in children and those immuno-compromised these disease is deadly and life altering.

Moreover, *Cryptosporidium* has been an organism that has been neglect over the years and also its intricate genome makes it difficult to study. Therefore, I plan to address these gaps by over the next five years by:

1. Annotating the *C. baileyi* genome sequence.
2. Comparing the *C. baileyi* genome sequence and annotation to other *Cryptosporidium* species.
3. Comparing the *C. baileyi* transcriptome and gene expression patterns to *C. parvum*.

In summary, the proposed research project consists of bioinformatics layered on top of strategic experimental data sets such as genome sequence and developmental RNA expression analysis. I used resources available through the UGA Georgia Advanced Computing Resource Center and trained an annotation pipeline called (Maker) by providing the unannotated genome sequence, annotated protein sequences from *C. parvum, C. hominis* and *C.muris* and RNA-seq data from *C. baileyi*. The result is a transcript and .gff file. Next, custom scripts will be written for determining how “correct” the annotation is using orthology, synteny, and general statistics. When the annotation is as “correct” as we can get it, then we can begin comparative genomics. The results of the comparative genomics will be seen through synteny that can be viewed in a genome browser of choice.

Types of Data

The annotation project will produce lots of data. The first data set produced will be the .gff file from the annotation from Maker. Custom scripts available through Maker software allow creation of data sets for intron (counts, size, location), exons (counts, size, location), and many other features (http://gmod.org/wiki/MAKER_Tutorial_2013).

Next, to test how “correct” the annotation, custom python and perl scripts will be written that will compare the predicted annotation set to other species using orthology and basic statistics.

The Standards to be used for Data and Metadata Format

The standards used for how I will format my data will follow that which is already available in CryptoDB. “The database, CryptoDB (http://CryptoDB.org), is a community bioinformatics resource for the AIDS-related apicomplexan-parasite, *Cryptosporidium*. CryptoDB integrates whole genome sequence and annotation with expressed sequence tag and genome survey sequence data and provides supplemental bioinformatics analyses and data-mining tools. A simple, yet comprehensive web interface is available for mining and visualizing the data,” (Heiges et. al, 2006). Moreover, I will use MAKER version 2.31.7. This software will produce files with the following format: (.gff) (.maker.proteins.fasta) (.maker.transcripts.fasta) and (run.log). The custom scripts will be in (.txt) (.pl) and (.py) format.

Policies for Access and Sharing

As the annotation reaches a point where we feel it is correct this data will be made public and deposited in CryptoDB and will be available for use and download by everyone in the science community. For use of the custom scripts I will ask that whoever interested emails the Kissinger Lab for permission.

Policies and Provisions for Re-use and Re-distribution

There are no restrictions on the re-use or re-distribution of the data because the data will be publically uploaded to CryptoDB. I just ask that the users cite the Kissinger Lab.

Plans for Archiving Data

The annotation was performed to create a long-term source to aid in the research of all *Cryptosporidium* species. More, CryptoDB is an NIH/NIAID-funded Bioinformatics Resource Center and the data when ready will be upload to the database and will be updated as need be to make this issue with *Cryptosporidium* an issue we can tackle as a community.

Use of Human Data

There will not be any human subjects used during the course of this research.

Additional data sharing requirements

Question not answered.