**BHB Genome Annotation Statistics**

<table>
<thead>
<tr>
<th></th>
<th>FTA</th>
<th>FTW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genome Length (bp)</strong></td>
<td>1,898,066 bp</td>
<td>1,898,476 bp</td>
</tr>
<tr>
<td>GC Content</td>
<td>32.1%</td>
<td>32.3%</td>
</tr>
<tr>
<td>Total Genes</td>
<td>1,897</td>
<td>1,897</td>
</tr>
<tr>
<td>Protein Coding Genes</td>
<td>1,580</td>
<td>1,634</td>
</tr>
<tr>
<td>Coding sequences</td>
<td>767</td>
<td>774</td>
</tr>
<tr>
<td>Genes annotated as miRNA</td>
<td>256</td>
<td>186</td>
</tr>
<tr>
<td>Structural RNAs</td>
<td>51</td>
<td>52</td>
</tr>
<tr>
<td>RNA Genes</td>
<td>50</td>
<td>58</td>
</tr>
<tr>
<td>Other RNA Genes</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>CDS annotated with complete Enzyme numbers (EC)</td>
<td>303</td>
<td>329</td>
</tr>
<tr>
<td>Taxonomy ID</td>
<td>458234</td>
<td>481136</td>
</tr>
<tr>
<td>RefSeq ID</td>
<td>NC_006949</td>
<td>NC_004253</td>
</tr>
<tr>
<td><strong>RefSeq ID</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Accession</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Assembly</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Species</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Strain</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Contig</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ABSTRACT**

The BioHealthBase (BHB) (www.biohealthbase.org) is one of eight Bioinformatics Resource Centers (BRC) funded by the National Institute of Allergy and Infectious Diseases (NIAID). The BHB serves as an integrated information resource for access to both biological data and analysis tools for five priority pathogens, including the bacterial pathogen *Francisella tularensis*. Curators at BHB have been actively involved in the expert annotation of several newly sequenced bacterial genomes including the manual curation of the automated predictions of gene structure, location and function. Our existing curation process involves gene prediction using the Glimmer algorithm based on the Interpolated Markov Model, automated annotation of predicted genes with the AutoAnnotate pipeline and manual curation of genes using the Manatee interface developed at the J. Craig Venter Institute (formerly TIGR). Manual curation includes editing of start and stop sites, assignment of functional features such as gene symbols, EC number and GO annotations followed by quality checks for data consistency and accuracy.


**INTRODUCTION**

The annotation of two complete genomes from the subtypes A and B has been completed by curators at BHB. The sequencing of *F. tularensis* subsp. *tularensis* RT96-JA1 and *F. tularensis* subsp. *holarctica* (FTA) was initially performed at the JGI production genomics (The Joint Genome Institute) facility in Walnut Creek, CA. Genome finishing was performed by the JGI at Los Alamos National Laboratory, Los Alamos, NM. The complete genomes were then auto-annotated at BHB for gene descriptions and functional predictions of proteins and all predictions were manually evaluated to provide reliable structural and functional assignments. Quality checks and data integrity evaluations were performed prior to submission of the annotated genomes to GenBank.

**METHODS**

**Fig.1** BioHealthBase Annotation Pipeline for *Francisella tularensis* Gene Prediction

- **Genome annotation** was performed at BHB using J. Craig Venter Institute’s (JCVI) Prokaryotic genome analysis pipeline for gene prediction (Glimmer3) automated annotation (AutoAnnotate) and manual curation (Manatee).
- The genomic sequence was first reformatted to ensure that dnaA was the initial gene and co-ordinates were reordered.
- The Glimmer3 (version 3.02) gene prediction algorithm which is based on Interpolated Markov Model was used with criteria imputed for initiation codon and specified length of ORF.
- The program RNAscansE (version 1.23) was employed to find RNAs. Exonumis (version 1.0.0) was used to identify 16S and 23S ribosomal RNA genes.
- RFam release 7.0, a comprehensive database of non-coding RNA (ncRNA) families was used to identify genes coding for other non-coding RNAs (such as 5S ribosomal RNAs).
- Prediction of ribosome binding sites (RBS) was done using RBSfinder algorithm developed by JCVI (former TIGR).

**Fig.2:** Annotation and Curation Pipeline

- **Proteins** were predicted on a genome against a non-redundant amino acid database (nr) that consists of all proteins available from GenBank, PIR and SWISS-PROT using the BLAST-Extend-Repraze (BER) algorithm and all significant matches were stored in a mini-database.
- A modified Smith-Waterman algorithm was performed on all the proteins in the mini-database.
- The gene was extended 300 nucleotides upstream and downstream of the predicted coding region to identify potential framewaeks.
- All protein sequences were searched against Pfam, HMMs and TIGRFAMs built from well-curated multiple alignments of proteins considered to share the same function or to be members of the same family.
- Prediction of transmembrane helices using TMHMM4, Signal peptide using signalP and lipoprotein motif and COG relationships of proteins based on phylogenetic classification of proteins encoded in complete genomes were also performed.
- The result data was sent to the AutoAnnotate pipeline at BHB to make functional predictions for proteins and were made available in the Manatee interface for manual evaluation of the AutoAnnotate predicted function.

**Manual Curation Process**

- JCVI’s guidelines for functional assignments were used for assignment of Gene symbol, EC number and GO annotations. Pseudogenes were termed ‘transposon-disrupted ORFs’ if they lacked experimental evidence.
- TIGRfam and Pfam HMM alignments along with Blast Extend Repraze (BER) alignments were used as primary evidence for functional annotations, and TMHMM, SignalP, COG relationships as well as literature data was used as supporting evidence. Enzyme Commission14 (EC) numbers were assigned to predicted enzymes based on their homology to an enzyme family.
- ORFs other than programmed frameshift or stop codons were reviewed for correct translation and when required, fixed by editing the start sites.

**REFERENCES**

11. SignalP using signalP
12. tRNAscan-SE
13. RFam
15. OrthoMCL: http://orthomcl.cbil.upenn.edu/cgi-bin/OrthoMcl

**ANALYSIS & RESULTS**

Using our completed genome annotation results, further analysis was performed at the Translational Genomics Research Institute, AZ, and a paper has been published on the ‘Complete Genomic Characterization of a Pathogenic A.III Strain of *Francisella tularensis* subspecies *tularensis*’ in *PLoS ONE*. (PMID: 17895988)

The analysis showed that FTW strain has some of the following features:

- Presence of 4,382 potential polymorphisms (3,367 SNPs and 1,015 indels)
- 31 large chromosomal rearrangements (inversions and translocations)
- One unique gene (FTW0088), hypothetical protein that is absent in Schu 4
- FTW is found to be 5.67 bp longer than Schu 4

**COMPARISON BETWEEN FTW AND SCHU4**

- **FEATURE**
  - **FTW**
  - **SCHU 4**
- **Transposons (Total)***
  - 82
  - 71
- **InFud (S630 family)**
  - 52 copies
  - 50 copies
- **InBta**
  - 18 copies
  - 16 copies
- **InSsh (SSNC family, InBta-152106)**
  - 3 copies
  - 2 copies, 1 fragment
- **InFud (S80283)**
  - 1 copy
  - 1 copy
- **InFud (S8043)**
  - 1 copy
  - 1 copy
- **InFud (S15959)**
  - 3 copies
  - 3 copies
- **ISod13**
  - 1
  - 1

Additionally, BioHealthBase also provides researchers with pre-computed Ortholog predictions for proteins based on OrthoMCL calculations, a list of all the IDEB curated epitopes as well as Protein 3-D structures for viewing and analysis, obtained from Protein Data Bank (PDB). Comparative Genomics tools and Systemy viewers will soon be available to perform whole genome analysis of the genomes in BHB.