University at Buffalo The State University of New York | REACHING OTHERS

SEPA SCIENCE EDUCATION PARTNERSHIP AWARD
Supported by the National Institutes of Health

Annotation of the *Legionella pneumophila* Genome from Locus Tags lpg0011-lpg0019

Abbigail Bartlett, Joseph Attili*, Megan Godfrey*, Raina Steger*, Julia Hundley and Caroline Preston

Greece Arcadia High School and the Western New York Genetics in Research and Health Care Partnership

Abstract

A group of consecutive 7 genes from the microorganism *Legionella Pneumophila (lpg0011-lpg0020)* were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene product name for 4 of these genes were assessed in terms of the general genomic information, amino acid sequence-based similarity data, structure-based evidence from the amino acid sequence, cellular localization data, and potential alternative open reading frames. The Genbank proposed gene product name did not differ significantly from the proposed gene annotation for 4 of the genes in the group and as such, the genes appear to be correctly annotated in the database. We did find some evidence for a better start codon of LPG 0019 that was 2 codons downstream of the computer called start codon..

Introduction

Legionella pneumopila is a species that is gram-negative bacilli. Also, it is a non-spore-forming gammaproteobacteria. It lies on the surface of water and is mostly found in showers, faucets, air-conditioners, fountains, public baths and many other places. It is not found commonly on natural waters, such as oceans, lakes or rivers. It is transmitted through human aerosols. Once in the aerosols, it multiplies and causes alveolar inflammation that is damaging to the body. Humans are considered incidental hosts, because legionella replicate within their alveolar macrophages. This may result as a febrile disease characterized by pneumonia. A sudden epidemic outbreak of legionella occurred between the attendees of the American Legion conference in Philadelphia. This outbreak uncovered 221 suspected cases of this respiratory disease where 34 individuals died. Legionella is responsible for 8,000 to 18,000 hospitalizations every year in the United States. The majority of cases occur in the summer. Those who smoke, have traveled recently, or have cancer, AIDS, or diabetes are more likely to be infected by this species. Legionella doesn't just affect humans. They are known to infect and replicate inside freshwater amoebae. Their association with amoebae may cause bacterial phenotypes (1).

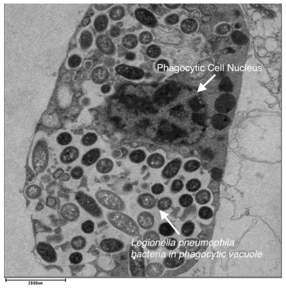


Figure 1. Transmission electron micrograph of *Legionella pneumophila* bacteria inside a phagocytic cell (2)

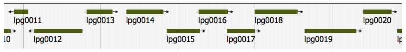


Figure 2. The arrangement of the locus tags of the *Legionella pneumophila* genome under investigation in this study

Methods and Materials

Modules of the GENI-ACT (http://www.geni-act.org/) were used to complete *Legionella pnemophila* genome annotation. The modules are described below:

Modules	Activities	Questions Investigated
Module 1- Basic Information Module	DNA Coordinates and Sequence, Protein Sequence	What is the sequence of my gene and protein? Where is it located in the genome?
Module 2- Sequence-Based Similarity Data	Blast, CDD, T-Coffee, WebLogo	Is my sequence similar to other sequences in Genbank?
Module 3- Cellular Localization Data	Gram Stain, TMHMM, SignalP, PSORT, Phobius	Is my protein in the cytoplasm, secreted or embedded in the membrane?
Module 4- Alternative Open Reading Frame	IMG Sequence Viewer For Alternate ORF Search	Has the amino acid sequence of my protein been called correctly by the computer?
Module 5- Structure-Based Evidence	TIGRfam, Pfam, PDB	Are there functional domains in my protein?

Results

Legionella Pneumophila 0011:

The initial proposed product of this gene by GENI-ACT was a Thiol-disulfide oxidoreductase. This gene product proposal was supported by the top BLAST hits for the amino acid sequence, the presence of well-curated protein functional domains within the amino acid sequence, the transmembrane topography of the amino acid sequence, and the cellular location of the amino acid sequence. As such, the proposed annotation is a Thiol-disulfide oxidoreductase.

Legionella Pneumophila 0015:

The initial proposed product of this gene by GENI-ACT was a multidrug resistance protein. This gene product proposal was supported by the top BLAST hits for the amino acid sequence, the presence of well-curated functional domains within the amino acid sequence, the cellular location of the amino acid sequence, and the enzymatic function of the amino acid sequence. As such, the proposed annotation is a multidrug resistance protein.

Legionella Pneumophila 0018:

The initial proposed product of these genes by GENI-ACT was a transporter protein. This gene product proposal was supported by the top BLAST hits for the amino acid sequences, the presence of well-curated functional domains within the amino acid sequences, the cellular location of the amino acid sequences, and the enzymatic function of the amino acid sequences. As such, the proposed annotation is a transporter protein.

Legionella Pneumophila 0019:

The initial proposed product of this gene by GENI-ACT was hemagglutinin. This gene product proposal was supported by BLAST hits from the amino acid sequence, functional domains within the amino acid sequence, the transmembrane topography of the amino acid sequence, the cellular location of the amino acid sequence. Inconsistencies and the alternate reading frame showed that the start codon is called incorrectly. I propose the new start codon begins at DNA coordinate 24117.

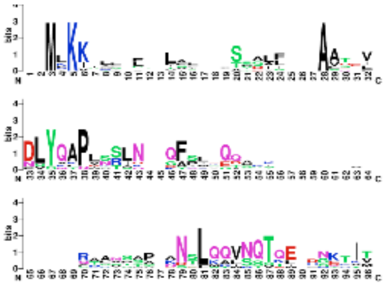
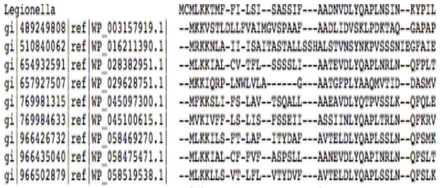


Figure 3 – *Legionella pnemophila 0019* Weblogo. Notice the gap in the first 2 positions and the poorly conserved regions through out.



```
Legionella                                  MCMLKKTMF-FI-LSI--SASSIF---AADNVDLYQAPLNSIN--KYPIL
gi|489249808|ref|WP_003157919.1|            --MKKVSTLDLLFVAIMGVSPAAF---AADLIDVSKLPDKTAQ--GAPAP
gi|510840062|ref|WP_016211390.1|            --MRKKNLA-II-ISAITASTALLSSHALSTVNSYNKPVSSSNIEGFAIE
gi|654932591|ref|WP_028382951.1|            --MLKKIAL-CV-TFL--SSSSLI---AATEVDLYQAPLNRLN--QFPLT
gi|657927507|ref|WP_029628751.1|            --MKKIQRP-LNWLVLA------G---AATGFPLYAAQMVTID--DASMV
gi|769981315|ref|WP_045097300.1|            --MFKKSLI-FS-LAV--TSQALL---AAEAVDLYQTPVSSLK--QFQLE
gi|769984633|ref|WP_045100615.1|            --MVKIVFF-LS-LIS--FSSEII---ASSIINLYQAPLTRLN--QFKRV
gi|966426732|ref|WP_058469270.1|            --MLKKILS-FT-LAF--ITYDAF---AVTELDLYQAPLSSLN--QFSMK
gi|966435040|ref|WP_058475471.1|            --MLKKIAL-CF-FVF--ASPSLL---AANEVDLYQAPINRLN--QFSLT
gi|966502879|ref|WP_058519538.1|            --MLKKLLS-VT-LFL--VTYDVF---AVTELDLYQAPLSSLN--QFSLK
```

Figure 4: *Legionella pneumophilia* T-Coffee Alignment. Notice, Legionella 0019 starts 2 amino acids sooner than the compared sequences.

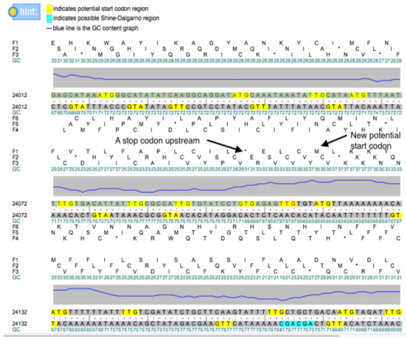
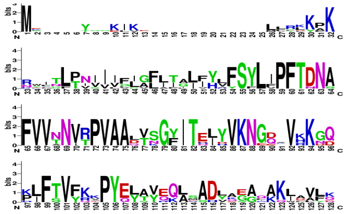
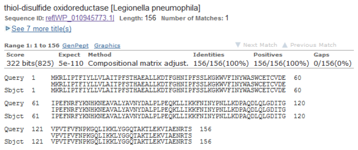


Figure 5 – *Legionella pneumophila* 0019 AORF. An ATG start codon downstream appears to fit more perfectly than the predicted codon



Legionella Pneumophila 0015 Weblogo. Near the beginning it is not well conserved but as you move towards the middle, it becomes more conserved.



thiol-disulfide oxidoreductase [Legionella pneumophila]
Sequence ID: ref|WP_010945773.1| Length: 156 Number of Matches: 1
See 7 more title(s)

Range 1: 1 to 156 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
322 bits(825)	5e-110	Compositional matrix adjust.	156/156(100%)	156/156(100%)	0/156(0%)

```
Query  1    MKRLIPTFIYLLVLAITPFFSTHAEALLKDTFGHNIPFSSLKGKWVFINYWASWCETCVDE  60
            MKRLIPTFIYLLVLAITPFFSTHAEALLKDTFGHNIPFSSLKGKWVFINYWASWCETCVDE
Sbjct  1    MKRLIPTFIYLLVLAITPFFSTHAEALLKDTFGHNIPFSSLKGKWVFINYWASWCETCVDE  60

Query  61   IPEFNRFYKNHKNEAVALYAVNYDALPLPEQKLLIKKFNINYPNLLKDPAQDLQLGDITG  120
            IPEFNRFYKNHKNEAVALYAVNYDALPLPEQKLLIKKFNINYPNLLKDPAQDLQLGDITG
Sbjct  61   IPEFNRFYKNHKNEAVALYAVNYDALPLPEQKLLIKKFNINYPNLLKDPAQDLQLGDITG  120

Query  121  VPVTFVFNPKGQLIKKLYGGQTAKTLEKVIAENRTS  156
            VPVTFVFNPKGQLIKKLYGGQTAKTLEKVIAENRTS
Sbjct  121  VPVTFVFNPKGQLIKKLYGGQTAKTLEKVIAENRTS  156
```

Figure 6. *Legionella pneumophila* 0011 Alignment of top hit. Notice the 100% identity and low E-Value score.

Conclusions

The GENI-ACT proposed gene product did not differ significantly for lpg0011, 0015 and 0018 from the proposed gene annotation for each of the genes in the group and as such, the genes appear to be correctly annotated by the computer database. Lpg0019 did not differ significantly from the proposed annotation, but the start codon did. This lead us to believe that the DNA coordinate of 24117 is better for the first nucleotide of the start codon.

Gene Locus	Geni-Act Gene Product	Proposed Annotation
LPG 0011	Thiol-disulfide oxidoreductase	Thiol-disulfide oxidoreductase
LPG 0015	Multidrug resistance protein	Multidrug resistance protein
LPG 0018	Transporter Protein	Transporter Protein
LPG 0019	Hemagglutinin	Hemagglutinin

References

1. Mercante and Winchell (2015) Current and Emerging *Legionella* Diagnostics for Laboratory and Outbreak Investigations Clin Microbiol Rev. 95–133.
2. https://upload.wikimedia.org/wikipedia/commons/f/fe/TEM_image_of_Legionella_pneumophila_within_a_phagocytic_cell.tif

Acknowledgments

Supported by an NIH Science Education Partnership (SEPA) Award - R25OD010536