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 *Haemophilus influenzae* Genome from Locus Tag HIBPF_00160 to HIBPF_00210

Jebadiah T. Braunscheidel*,Evan Fanara*, Geyenna Sterling-Lentsch*, Tabitha Raithel*, Michael Rechichi*, John Velasquez*, and Jay Rodemeyer

West Seneca West Senior High School 3330 Seneca Street, West Seneca, NY 14224 and The Western New York Genetics in Research Partnership

*indicates equal contribution by gene annotation authors

Abstract

A group of six consecutive genes from the microorganism *Haemophilus influenzae* (HIBPF_00160 - HIBPF_00210) were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene product names for each gene and assessed them in terms of general genomic information, amino acid sequence-based similarity data, structure-based evidence from the amino acid sequence, cellular localization data, potential, or absence of gene duplication and degradation, and enzymatic function. The Genbank proposed gene product names did not differ significantly from the proposed gene annotations for all six of the genes in the group, which, appear to be correctly annotated by the database.

Introduction

Haemophilus influenzae is a Gram-negative, coccobacillary, facultatively anaerobic pathogenic bacterium belonging to the *Pasteurellaceae* family. *H. influenzae* was first described in 1892 by Richard Pfeiffer during an influenza pandemic.

The bacterium was mistakenly considered to be the cause of influenza until 1933 when the viral cause of influenza became apparent, and is still colloquially known as 'bacterial influenza'. *H. influenzae* is responsible for a wide range of localized and invasive infections. This species was the first free-living organism to have its entire genome sequenced by Craig Venter and his team at The Institute for Genomic Research.

Most strains of *H. influenzae* are opportunistic pathogens; that is, they usually live in their host without causing disease, but cause problems only when other factors (such as a viral infection, reduced immune function or chronically inflamed tissues, e.g. from allergies) create an opportunity. They infect the host by sticking to the host cell using trimeric autotransporter adhesins.

Naturally acquired disease caused by *H. influenzae* seems to occur in humans only. In infants and young children, *H. influenzae* type b (Hib) causes bacteremia, pneumonia, epiglottitis and acute bacterial meningitis.

The genome of *H. influenzae* consists of 1,830,140 base pairs of DNA in a single circular chromosome that contains 1740 protein-coding genes, 2 transfer RNA genes, and 18 other RNA genes. The sequencing method used was whole-genome shotgun, which was completed and published in *Science* in 1995.

Students were given a gene sequence to annotate and compare with Genebank proposed annotations. Specific information was collected from NCBI BLAST, T-Coffee, WebLogo, TIGRFAM, PFam, PDB, HMM Logos, Protein Data Bank (PDB), TMHMM, Signal IP, PSORTb, Phobius Applications, and the IMG database to conclude if their findings agree with the proposed annotation.

Figure 1: A scanning electron micrograph of *Haemophilus influenzae* showing the characteristic growth pattern of the individual bacterial cells and typical conjunctivitis in a six month old child.

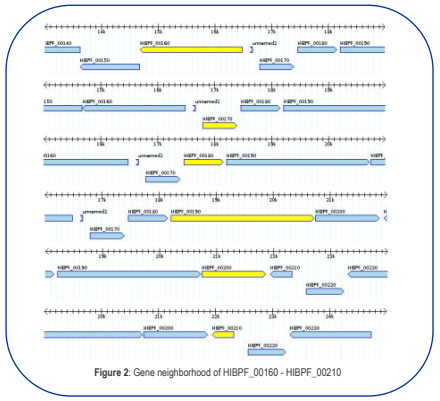

Figure 2: Gene neighborhood of HIBPF_00160 - HIBPF_00210

Methods and Materials

Modules of the GENI-ACT (http://www.geni-act.org/) were used to complete *Haemophilus influenzae* 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 - Structure-Based Evidence	TIGRfam, Pfam, PDB	Are there functional domains in my protein?
Module 4 - Cellular Localization Data	Gram Stain, TMHMM, SignalIP, PSORTb, Phobius	Is my protein in the cytoplasm, secreted or embedded in the membrane?
Module 5 - 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 8 - Evidence for Horizontal Gene Transfer	Phylogenetic Tree	Has my gene co-evolved with other genes in the genome?
Final Annotation	Review data from all modules	Does the student proposed name of the gene agree with that proposed by the automated computer annotation? Are any changes proposed to the pipeline annotation?

Results

HIBPF_00160:

Initial proposed results for HIBPF_00160 Using cellular localization data was for GTP-binding membrane protein LepA, which was determined by the Genebank database and is cytoplasmic in nature.

This conclusion was reached using data from TMHMM, LipoP, and signalP suggesting that the HIBPF_00160 resides in the cytoplasm due to no predicted cleavage sites or transmembrane helices and is potentially responsible for gene regulation, ribosome assembly, transport and/or respiration.

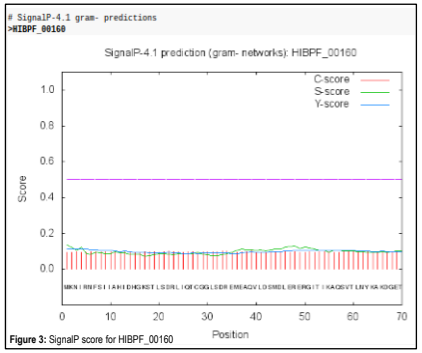

Figure 3: SignalP score for HIBPF_00160

HIBPF_00170:

Using structural based evidence, key functional and structural residues of the HMM Logo, along with pairwise alignment, suggest that HIBPF_00170 is a fimbrial protein used to adhere to cells for infection.

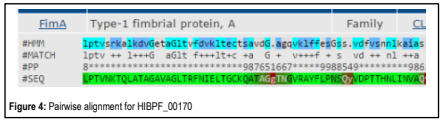

Figure 4: Pairwise alignment for HIBPF_00170

HIBPF_00180:

Acording to PSORT, the protein can be found in the periplasm. Data acquired from Phobius suggests the protein functions as a potential chaperone. The LipoP prediction in the figure below shows significant results for a cleavage site in the cellular membrane.

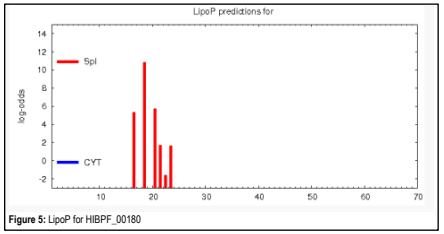

Figure 5: LipoP for HIBPF_00180

HIBPF_00190:

The initial proposed product of this gene by GENI-ACT was a fimbrial usher protein. This gene product proposal was supported by results gathered by using sequence-based similarity data.

WebLogo for the amino acid sequence show significant areas of high conservation, specifically in areas 339-312 and 353-359. This suggests that this protein is similar to fimbrial adhesion proteins suggested by the Genebank, and aids in binding to its host.

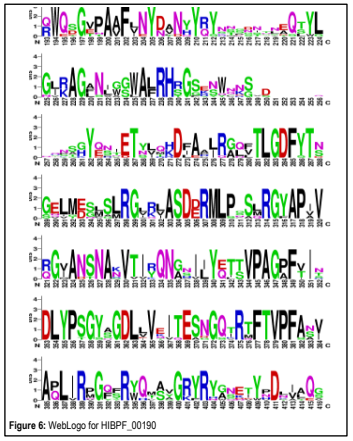

Figure 6: WebLogo for HIBPF_00190

HIBPF_00200:

The proposed product of this gene by GENI-ACT is a fimbrial adhesin protein. Figure 7, right, depicts the Crystal structure of F17b-G in complex with N-acetyl-D-glucosamine found in *Haemophilus influenzae*. This protein has been recognized as the most important variable in bacterial crystallization.

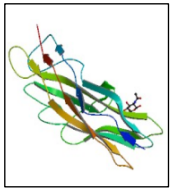

Figure 7: Crystal structure of F17b-G in complex with N-acetyl-D-glucosamine in *H. influnzae*

HIBPF_00210:

Phylogeny of proposed hypothetical acid-induced glycylradical enzyme by Genebank was supported by data acquired through research of horizontal gene transfer of *H. Influenzae*.

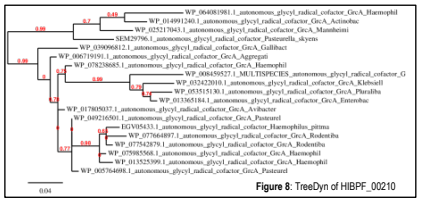

Figure 8: TreeDyn of HIBPF_00210

Conclusion

The products proposed by GENI-ACT did not differ significantly from the proposed gene annotation for all six of the genes in the group and as such, those genes appear to be correctly annotated by the computer database.

Gene Locus	Geni-Act Products	Proposed Annotation
HIBPF_00160	GTP-binding membrane protein LepA	GTP-binding membrane protein LepA
HIBPF_00170	Fimbrial protein	Fimbrial protein
HIBPF_00180	putative fimbrial chaperone protein	putative fimbrial chaperone protein
HIBPF_00190	fimbrial usher protein	fimbrial usher protein
HIBPF_00200	Fimbrial adhesin	Fimbrial adhesin
HIBPF_00210	conserved hypothetical acid-induced glycylradical enzyme	glycyl radical cofactor GrcA

References

Hallström, T., & Riesbeck, K. (2010, June). Haemophilus influenzae and the complement system. Retrieved Dec. 12, 2016, from https://www.ncbi.nlm.nih.gov/pubmed/20399102

Acknowledgments

Special thank you to SUNY at Buffalo, Dr. Stephen Koury, and Dr. Rama Dey-Rao for the opportunity and their assistance with this project.

Supported by an NIH Science Education Partnership (SEPA) Award - R25OD10536-1 A1