Annotation of a CRISPR array from the *Nanoarchaeum equitans Kin4-M* Genome (Locus Tags NEQ019, NEQ021 and NEQ022)

Quinn Hartman, Vieve Sherwood, Catherine Liu and Peter Hentschke
The Harley School, 1981 Clover St. Rochester NY, 14618 and the Western New York Genetics in Research and Health Care Partnership

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

Abstract

A group of three genes associated with a CRISPR array from the recently-discovered hydrothermal vent Archaea called *Nanoarchaeum equitans Kin4-M* was annotated using the collaborative genome annotation website GENI-ACT. The gene product names were not predicted by GENI-ACT. However, in this research the genes were analyzed in terms of their general genomic information, amino acid sequence-based similarity data, structure-based evidence from the amino acid sequence, and cellular localization data. Based on these results, proposed gene product annotations are provided.

Introduction

Nanoarchaeum equitans is a marine Archaea species. It was found in a hydrothermal vent off the coast of Iceland for the first time in 2002. Many scientists are considering making it the first species in a brand new phylum. *Nanoarchaeum equitans* and *Ignicoccus hospitalis* have an extremely unique relationship. These are two archaeal species that rely on each other. *Nanoarchaeum equitans* has a very reduced genome, and therefore, must get metabolites and energy from the *Ignicoccus hospitalis* by attaching to it. Also, a similarity in the lipid compositions of the two species lead scientist to believe there are possibly transport mechanisms between the two species. (Podar et. al., 2008).

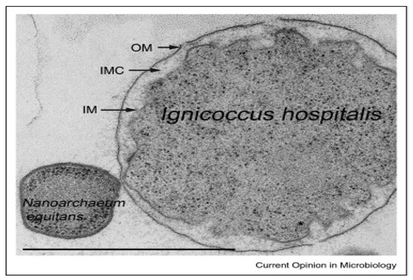

Figure 1. A picture of *Nanoarchaeum Equitans* and its host *Ignicoccus Hospitalis.* (Moissl-Eichinger & Huber, 2011).

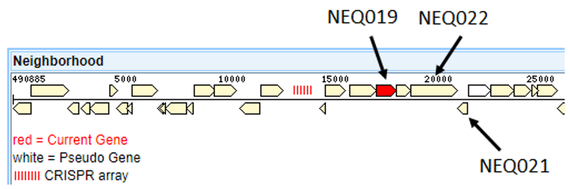

Figure 2. The locus tags and relative position of the genes under investigation in this research

Methods

Modules of GENI-ACT (http://www.geni-act.org/) were used to annotate genes from the *Nanoarchaeum equitans Kin4-M* genome. The modules used are described below:

Modules	Activities	Questions Investigated
Basic Information	DNA Coordinates and Sequence, Protein Sequence	What is the sequence of the gene and protein? Where is it located in the genome?
Sequence-Based Similarity	Blast, CDD, T-Coffee, WebLogo	How similar is the protein under investigation to other proteins in GenBank?
Structure-Based Similarity	TIGRfam, Pfam, PDB	What functional domains are present in the protein under investigation?
Cellular Localization	Gram Stain, TMHMM, SignalP, LipoP, Psortb, Phobius	Is the protein under investigation located in the cytoplasm, secreted, in the periplasm or embedded in the cell membrane or cell wall?
Final Annotation	Evaluate data from all modules	Has the gene been correctly called by the pipeline annotation?

Results

NEQ019:
NEQ019 refers to a gene located at the coordinates 17628..18596. NEQ019 has only 3 hits on BLAST using SwisProt, all of which have scores that are <40 indicating poor alignments. The top two NR BLAST alignments indicated the gene sequence most closely matched a type I-B CRISPR-associated protein.

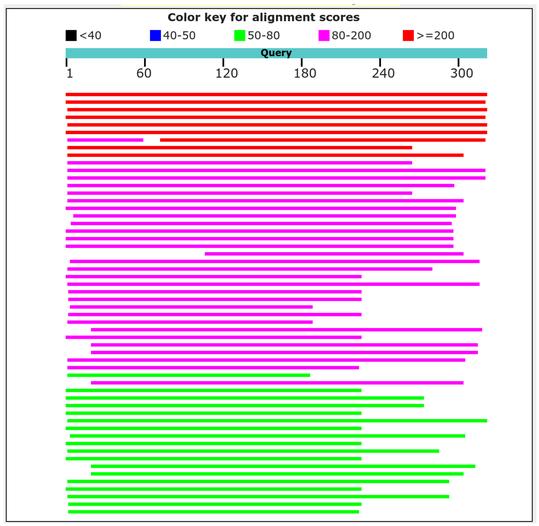

Figure 3. The BLAST (NR) alignments with NEQ019.

The NR BLAST color alignments graph for NEQ019 showed many hits with a consistent shorter alignment in the vicinity of residue 220. This might mean that there is a conserved domain within the sequence.

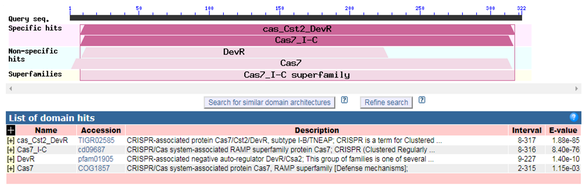

Figure 4. The CDD hits for NEQ019 identifying conserved domain similarities with the CRISPR-associated proteins in TIGRFAM and Pfam.

NEQ021:
The gene NEQ021 is found at coordinates 21555 to 22061. The protein sequence has 168 amino acids and the nucleotide sequence has 507 bases. According to TIGRFAM, the gene is predicted to be a CRISPR-associated protein Cas4 which represents a family of proteins associated with CRISPR repeats in a wide set of prokaryotic genomes and the function of this protein is undefined. Pfam results show that the domain of the gene contains three conserved cysteines at its C terminus end and belong to the clan PDDEXK. TMHMM, SignalP and Phobius results found no transmembrane helices or signal peptides, however PSORT-B states there is not enough evidence to identify the protein's location.

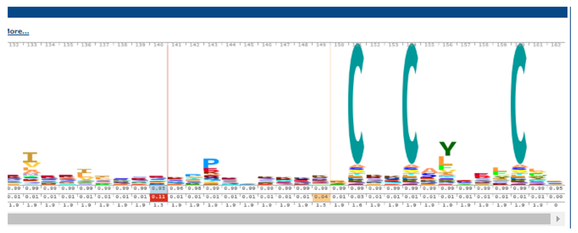

Figure 5. The carboxyl terminal end of the HMM logo from the Pfam of NEQ021. This shows the protein domain of the gene which contains three conserved cysteines at its C terminus.

NEQ022:
This gene has the coordinates 19294 to 21561 as predicted by GENI-ACT. The nucleotide sequence is 2268 bases long while the 755 amino acid protein sequence. A TIGRFAM hit predicts the gene to be a CRISPR-associated helicase Cas3, meaning it is part of an immune system that gives protection against viruses and other mobile genetic elements. The PDB database has a good match to this gene that is identified as a CRISPR-associated protein (Code 4Q2C). TMHMM predicts zero transmembrane helices in the gene and the PSORTB final prediction supports the conclusion that it is located in the cytoplasm. This location is consistent with a CRISPR-associated protein.

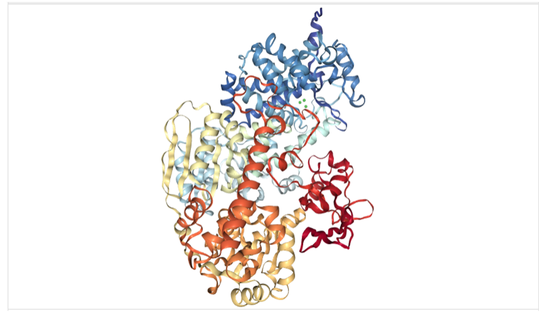

Figure 6. PDB Crystal structure of NEQ022 which is identified as CRISPR-associated protein.

Conclusion

From the annotation analysis done during the period of the project, the gene products proposed in the research are shown in the table below.

Locus Tag	Proposed Annotation
NEQ019	CRISPR-associated autoregulator
NEQ021	CRISPR-associated exonuclease Cas4 family
NEQ022	CRISPR-associated helicase

References

Moissl-Eichinger & Huber, 2011. Archaeal symbionts and parasites. Current opinion in microbiology, 14(3): 364-70.

Podar et al., 2008. A genomic analysis of the archaeal system *Ignicoccus hospitalis-Nanoarchaeum equitans*. Genome Biol. 9(11):R158.

University at Buffalo The State University of New York
Jacobs School of Medicine and Biomedical Sciences
buffalo.edu

Acknowledgments

Supported by an NIH Science Education Partnership (SEPA) Award - R25GM129209