Comparing the annotation abilities of three annotation databases using the *Halorhabdus utahensis* genome

Pallavi Penumetcha[1], Peter Bakke[1], Nick Carney[1], Will DeLoache[1], Mary Gearing[1], Matt Lotz[1], Jay McNair[1], Samantha Simpson[1], Max Win[1], Laura Voss[1], Laurie Heyer[2] and Malcolm Campbell[1].

1 Department of Biology, Davidson College, Davidson, North Carolina, United States of America, **2** Department of Mathematics, Davidson College, Davidson, North Carolina, United States of America

Abstract:

The purpose of this study was to compare the annotation abilities of three different databases: Joint Genome Institute (JGI), Rapid Annotation using Subsystems Technology server (RAST) and the J. Craig Venter Institute's annotation database: Manatee. Understanding the strengths and weaknesses of certain annotation tools can be very beneficial to gaining a more complete understanding of a novel organism's genome. Specifically, we chose to look at the genome of the archaeal halobacterium *Halorhabdus utahensis*. Studying archaeal enzymes can be important for developing novel biotechnological processes. For example, halophilic archaea have been found to be useful in degrading organic pollutants that result from textile production (3). In order to gain a clear understanding of each ORF studied, we used online tools such as BLAST, CDD, Pfam and KEGG Pathway to validate the ORF calls made by the annotation databases. We found that there was a lot of disagreement in the ORF and protein name calls made by the three databases. There were also discrepancies in the EC numbers associated with certain protein names and amino acid sequences. These data showed us that manual investigation is required to gain an accurate picture of particular ORFs in a novel organism. At the termination of this study, we determined that it is difficult to conclude which website did a better job of annotating the genome, because they all had characteristics that made them useful in determining the overall metabolic pathways this organism has. The methods used in this study can be utilized to gain more accurate annotations of genomes in novel organisms. In the future, we hope to extend this study and complete this three-way comparison of the *H. utahensis* genome in order to gain a better understanding of how these three databases differ in their annotation abilities.

Introduction:

Annotating genomes of novel organisms is a very delicate and time consuming process. Many times, scientists rely solely on annotation databases to provide complete sequence information about these novel genomes, which can be very detrimental to the correct annotation of the genome if there is any mistake in how a computer program calls open reading frames (ORFs.) The purpose of this study was to compare the annotations of three different databases in order to see how accurately annotation databases make ORF calls. We decided to investigate the genome of a halobacterium called *Halorhabdus utahensis*. There has been an increasing interest in studying organisms from extreme environments, because these organisms have very unique metabolic processes that allow them to survive. Basic research with these organisms can give insight into how these enzymes function in such extreme environments, which can be beneficial to the development of novel biotechnological processes (3). For example halophilic archaea are very useful in degrading organic pollutants, which result from textile production (3). Halophiles are also very useful in their ability to produce biopolymers, which is particularly important, because these are compounds that chemists have trouble producing synthetically (1). These examples show that studying the genome of a novel organism can be beneficial to finding cheaper and more environmentally friendly biotechnological processes.

Halorhabdus utahensis is a halophilic archaea that was isolated from the Great Salt Lake. It is an extremophile and grows optimally in an environment that is 27% (w/v) NaCl, which appears to be the highest reported salinity optimum for any living organism (4). This organism has many characteristics that categorize it as a halobacterium, but also properties that make it a unique organism in the halobacterium genus. For example, the major lipids present in this organism make it a part of the halobacterium genus, but this organism's 16S rRNA sequence and inability to use complex substrates for growth make it a distinctive member of this genus (4). General characteristics of this organism's genome have been studied, but an in depth analysis of the proteins and metabolic pathways in this organism have never been closely examined. In this study, we compared the gene and pathway annotations of this organism in three different databases: the Joint Genome Institute (JGI), Rapid Annotation using Subsystems Technology server (RAST) and the J. Craig Venter Institute's annotation database: Manatee. We believe that this study will not only give a better understanding of this organism, but will also illuminate the advantages and disadvantages of using online databases to study an organism's genome. Such a comparison has never been undertaken and we believe the information gained from this study will be very useful in understanding the importance of further investigating ORF calls made by online databases.

Materials and Methods:

In order to annotate the genes and pathways in *H. utahensis*, we used a variety of annotation databases and online tools. Our primary source of sequence information came from the three annotation databases: JGI, RAST and Manatee. All three databases gave both DNA and protein sequences for all ORFs called, as well as other general information about the *H. utahensis* genome, such as the number of tRNA genes. In addition to this information, RAST provided KEGG pathways that had EC numbers associated with enzymes in our organism highlighted. In addition to the information we received from the annotation databases, we used other online tools to compare and/or confirm the information we received from the three databases. The most commonly used online tools were NCBI BLAST, BLAST 2, Conserved Domains Database (CDD), Pfam, KEGG Pathway and ExPASy Enzyme Nomenclature Database. NCBI BLAST was used to compare the protein sequences found in our organism against known protein sequences in other organisms. BLAST2 was useful in comparing amino acid sequences between the different databases and in comparing amino acid sequences in the databases to amino acid sequences associated with particular EC numbers. CDD and Pfam were used to confirm that a protein sequence had functional domains that allowed it to have a certain function. We used KEGG Pathway to see which enzymes were present in organisms closely related to *H. utahensis.* In this database, I looked at the metabolic pathways in *Halobacterium salinarum*, because this is the halobacterium that is most closely related to our organism. I used the ExPASy Enzyme Nomenclature Database to find the names associated with specific EC numbers and to find alternative names for certain enzymes. Searching for alternative enzyme names was very important in confirming whether or not an enzyme was present in our organism.

In addition to these pre-existing online tools, I also used tools developed by members of our class. One of the most frequently used tools performs a text-based search between the three databases and retrieves the protein sequences, protein name and EC number (if available) associated with a particular text. This tool was very useful in comparing protein calls between different databases and determining whether an enzyme was present in a pathway that was not highlighted in the RAST KEGG pathway.

The methods used in this project were largely created throughout the course of the class. However, towards the end of the process, I was able to create a more standardized method for annotating specific genes. The following flow diagram gives a basic outline of the methods I used to further investigate gene calls made by the annotation databases.

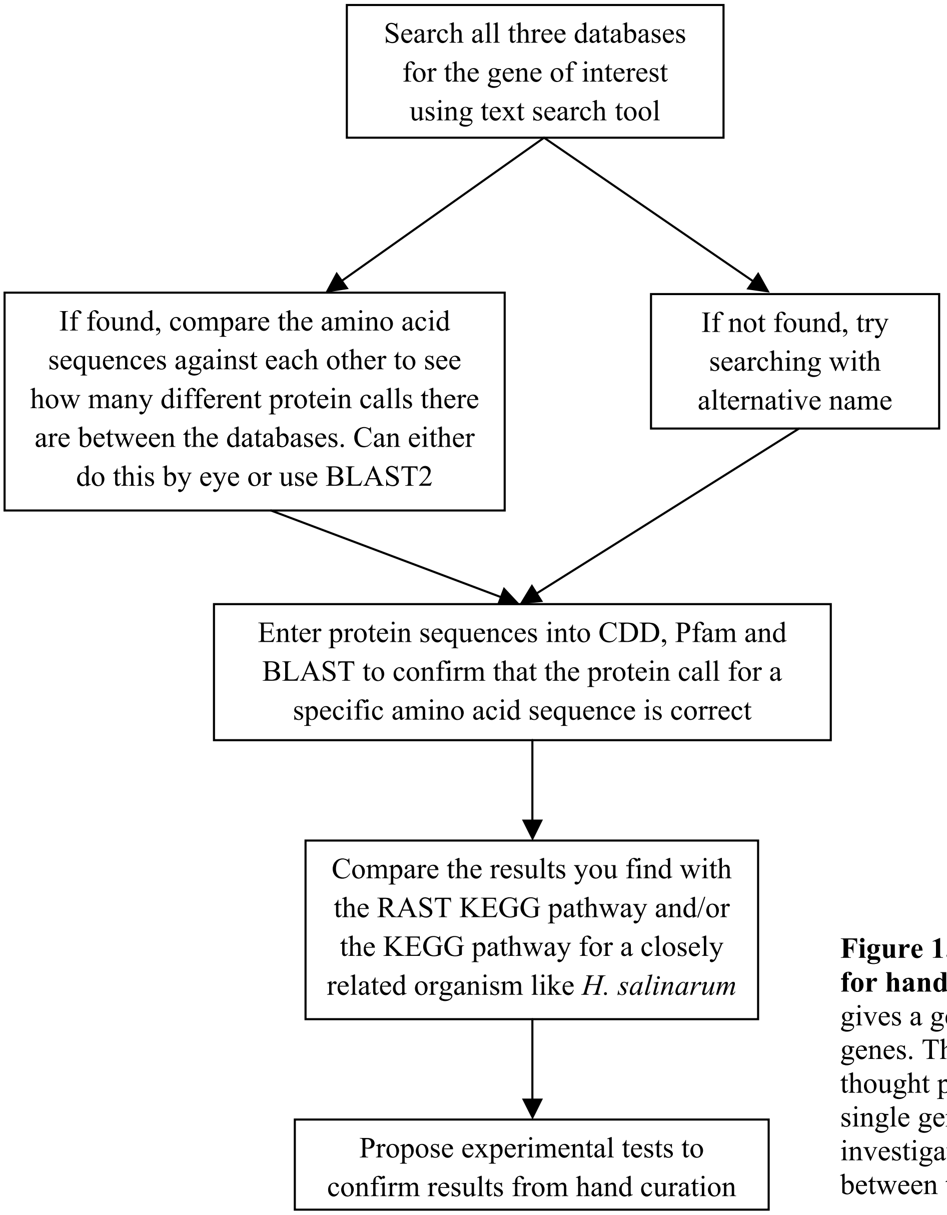

Search all three databases for the gene of interest using text search tool
If found, compare the amino acid sequences against each other to see how many different protein calls there are between the databases. Can either do this by eye or use BLAST2
If not found, try searching with alternative name
Enter protein sequences into CDD, Pfam and BLAST to confirm that the protein call for a specific amino acid sequence is correct
Compare the results you find with the RAST KEGG pathway and/or the KEGG pathway for a closely related organism like *H. salinarum*
Propose experimental tests to confirm results from hand curation

Figure 1. Flow diagram of general methods for hand curation of genes. This diagram gives a general idea of how I investigated genes. This does not necessarily show my thought process for the hand curation of every single gene, but gives a general sense of how to investigate an ORF that has discrepancies between the three databases.

In order to keep a record of the progress we made, we kept an online "lab notebook," in the form of a wiki, as a forum to post new information and questions. This lab notebook contains tutorials on how to use many of the most commonly used online tools under "Tutorials for Annotating Genomes." We also downloaded the nucleotide and protein sequences from the three databases into FASTA format, so that they were easily accessible. These files can be found at the beginning of the lab notebook under "Links to Multiple Databases." This website also contains in depth information about specific genes and pathways that members of the class investigated and a list of glossary words that are common to field of genomics. The lab notebook can be accessed at http://gcat.davidson.edu/GcatWiki/index.php/Halorhabdus_utahensis_Genome.

Results

The most striking result of our research is that the three databases delivered drastically different information. There were major discrepancies in ORF calls and in how particular ORFs were named. For example, JGI called 3076 ORFs, RAST called 2867 ORFs and Manatee call 3238 ORFs. In order to investigate why the databases were not calling the same ORFs, we created a pairwise comparison of all three databases based on the start number of the gene and we also investigated which start codons the different databases used. With the pairwise comparison we found that only 1471 genes, between all three databases, matched at both the start and stop sites.

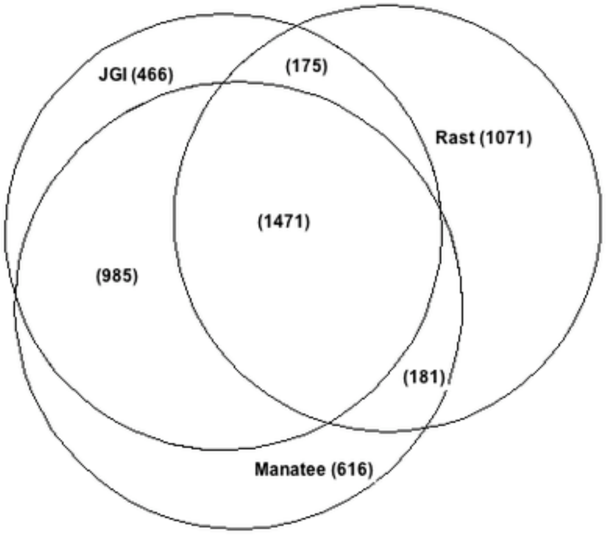


Figure 2. Venn diagram comparing the number of start and stop site matches in the three database. This figure gives an example of how differently the three databases annotated this genome. This diagram suggests that the three databases probably used different start codons in calling ORFs, which may have led to such different ORF calls between the three databases.

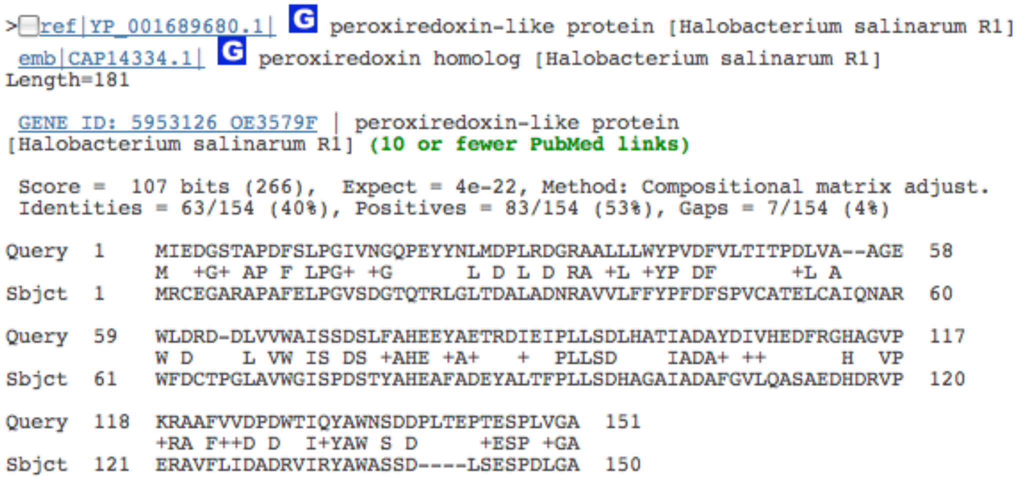
Because of this discrepancy in start and stop site matches, we decided to further investigate the start codons utilized by the different databases. We found that RAST was more likely to use start codons other than ATG in calling an ORF and that JGI was the least likely to use a start codon other than ATG. This discrepancy in start codons used by the databases suggests that there could be differences in how the databases named a particular ORF.

Start Codon	JGI Predictions	RAST Predictions	Manatee Predictions
ATG	2604	1723	2562
Other	493	1175	692
Total	3097	2898	3254
Percentage Not ATG	15.9%	40.5%	21.3%

Note: 94% of the alternatives start codons were TTG or GTG. None were CTG.

Table 1. Start codons utilized by the three databases. This table shows that the three databases use differing start codons with varying degrees of frequency to call ORFs. This probably contributes to the lack of matches in the start and stop sites between the three databases.

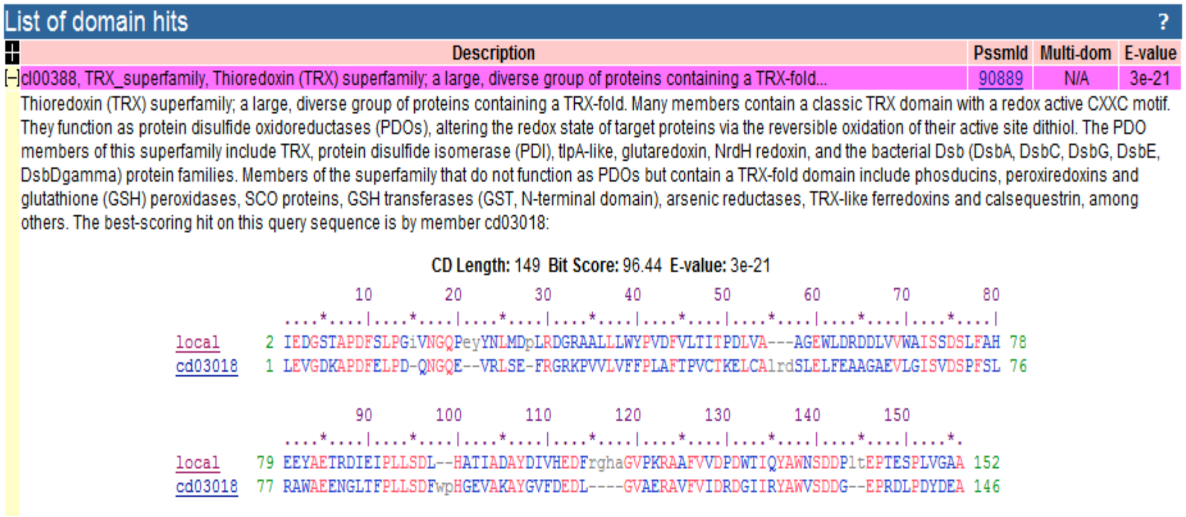
We decided to study individual genes to investigate such discrepancies in calling ORFs. The example I focus on here is a protein call that JGI and Manatee called a peroxiredoxin (EC number 1.11.1.15-an enzyme that purges peroxide from a cell) and that RAST called a monooxygenase (EC number 1.14.13-an enzyme that adds a single atom of oxygen to its substrate). This gene has a mismatched ORF start number and the RAST protein sequence is 7 amino acids longer than the JGI sequence. The start codon in the RAST sequence is GTG and the start codon in the JGI sequence is ATG. It was important to see whether or not this 7 base pair discrepancy conferred a different function on the ORF called in the RAST database. The first step was to BLAST the JGI and RAST protein sequences in NCBI's BLASTp. For both of the protein sequences, the top hits were monooxygenases, which would suggest that this amino acid sequence had functional domains that made it a monooxygenase. However, the hits for the JGI protein sequence were interesting, because the monooxygenase hit was followed by two peroxiredoxins, one of which came from *Halobacterium salinarum*, which is the halobacterium that is most closely related to *H. utahensis.*



```
>ref|YP_001689680.1| G peroxiredoxin-like protein [Halobacterium salinarum R1]
 emb|CAP14334.1| G peroxiredoxin homolog [Halobacterium salinarum R1]
Length=181

 GENE ID: 5953126 OE3579F | peroxiredoxin-like protein
[Halobacterium salinarum R1] (10 or fewer PubMed links)

 Score =  107 bits (266),  Expect = 4e-22, Method: Compositional matrix adjust.
 Identities = 63/154 (40%), Positives = 83/154 (53%), Gaps = 7/154 (4%)

Query  1    MIEDGSTAPDFSLPGIVNGQPEYYNLMDPLRDGRAALLLWYPVDFVLTITPDLVA--AGE  58
            M  +G+ AP F LPG+ +G       L D L D RA +L +YP DF       +L A
Sbjct  1    MRCEGARAPAFELPGVSDGTQTRLGLTDALADNRAVVLFFYPFDFSPVCATELCAIQNAR  60

Query  59   WLDRD-DLVVWAISSDSLFAHEEYAETRDIEIPLLSDLHATIADAYDIVHEDFRGHAGVP  117
            W D    L VW IS DS +AHE +A+    +  PLLSD    IADA+ ++       H  VP
Sbjct  61   WFDCTPGLAVWGISPDSTYAHEAFADEYALTFPLLSDHAGAIADAFGVLQASAEDHDRVP  120

Query  118  KRAAFVVDPDWTIQYAWNSDDPLTEPTESPLVGA  151
            +RA F++D D  I+YAW S D      +ESP +GA
Sbjct  121  ERAVFLIDADRVIRYAWASSD----LSESPDLGA  150
```

Figure 3. BLAST alignment of ORF 109611..110123 in JGI. This alignment gives evidence for this ORF being a peroxiredoxin, no only because the alignment comes from a closely related halobacterium, but also because the e-value (4e-22) is very small.

I thought that this result required more investigation, because it came from an organism that is so closely related to *H. utahensis.* I entered the amino acid sequences from RAST and JGI into CDD and both sequences returned the exact same functional domains. In addition, the superfamily that was detected in this protein sequence, called thioredoxin, is a superfamily that is related to peroxiredoxins.



List of domain hits ?

Description	PssmId	Multi-dom	E-value
cl00388, TRX_superfamily, Thioredoxin (TRX) superfamily; a large, diverse group of proteins containing a TRX-fold...	90889	N/A	3e-21

Thioredoxin (TRX) superfamily; a large, diverse group of proteins containing a TRX-fold. Many members contain a classic TRX domain with a redox active CXXC motif. They function as protein disulfide oxidoreductases (PDOs), altering the redox state of target proteins via the reversible oxidation of their active site dithiol. The PDO members of this superfamily include TRX, protein disulfide isomerase (PDI), tlpA-like, glutaredoxin, NrdH redoxin, and the bacterial Dsb (DsbA, DsbC, DsbG, DsbE, DsbDgamma) protein families. Members of the superfamily that do not function as PDOs but contain a TRX-fold domain include phosducins, peroxiredoxins and glutathione (GSH) peroxidases, SCO proteins, GSH transferases (GST, N-terminal domain), arsenic reductases, TRX-like ferredoxins and calsequestrin, among others. The best-scoring hit on this query sequence is by member cd03018:

CD Length: 149 Bit Score: 96.44 E-value: 3e-21

```
                  10        20        30        40        50        60        70        80
          ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local    2 IEDGSTAPDFSLPGiVNGQPeyYNLMDpLRDGRAALLLWYPVDFVLTITPDLVA---AGEWLDRDDLVVWAISSDSLFAH 78
cd03018  1 LEVGDKAPDFELPD-QNGQE--VRLSE-FRGRKPVVLVFFPLAFTPVCTKELCAlrdSLELFEAAGAEVLGISVDSPFSL 76

                  90       100       110       120       130       140       150
          ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*.
local   79 EEYAETRDIEIPLLSDL--HATIADAYDIVHEDFrghaGVPKRAAFVVDPDWTIQYAWNSDDPltEPTESPLVGAA 152
cd03018 77 RAWAEENGLTFPLLSDFwpHGEVAKAYGVFDEDL----GVAERAVFVIDRDGIIRYAWVSDDG--EPRDLPDYDEA 146
```

Figure 4. Results of JGI sequence entered in CDD (results are the same for RAST sequence). These results give strong evidence that this amino acid sequenc is a peroxiredoxin, both because the superfamily is related to peroxiredoxins anc because of the low e-value (3e-21) for the alignment.

The results from CDD provide compelling evidence for this gene being a peroxiredoxin, because this database showed that a protein sequence that was named a monooxygenase returned a protein family associated with peroxiredoxins. This evidence also gives us insight into the fact that these three databases are clearly using different techniques to name proteins. This idea is further illustrated in a Venn diagram that shows how differently the three databases called peroxiredoxins. In order to obtain this information, I searched all three databases for proteins that were named peroxiredoxins, retrieved the protein sequences and compared them against each other.

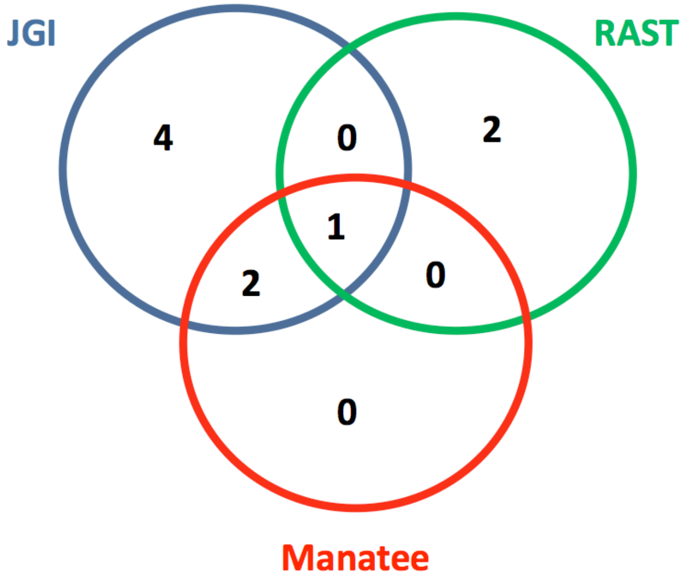


Figure 5. Venn diagram comparing amino acid sequence matches for peroxiredoxin protein calls. This figure was generated by comparing amino acid sequences that were associated with the protein name peroxiredoxin in the three databases. This figure shows that, for a single protein, there is a great deal disagreement among the databases as to which protein sequences are associated with this protein name.

This investigation of single ORF shows how much variation there is in how the three databases annotate a single gene. This investigation also gives us a better idea of how to validate the protein calls made by the three databases. The next step was to investigate the discrepancies in pathway annotation between the three databases. Investigating an organism's metabolic pathways allows us to see what unique properties this organism may have. This was more complicated than looking at genes, because the pathway maps are very inter-connected. In order to focus on the main ideas the pathway presents, I hand annotated the pathway I was interested in and color-coded the enzymes based on whether or not they were highlighted by the RAST annotation and whether or not they were present in the other databases. The following image is a hand curation of the glycolysis/gluconeogenesis pathway found in RAST.

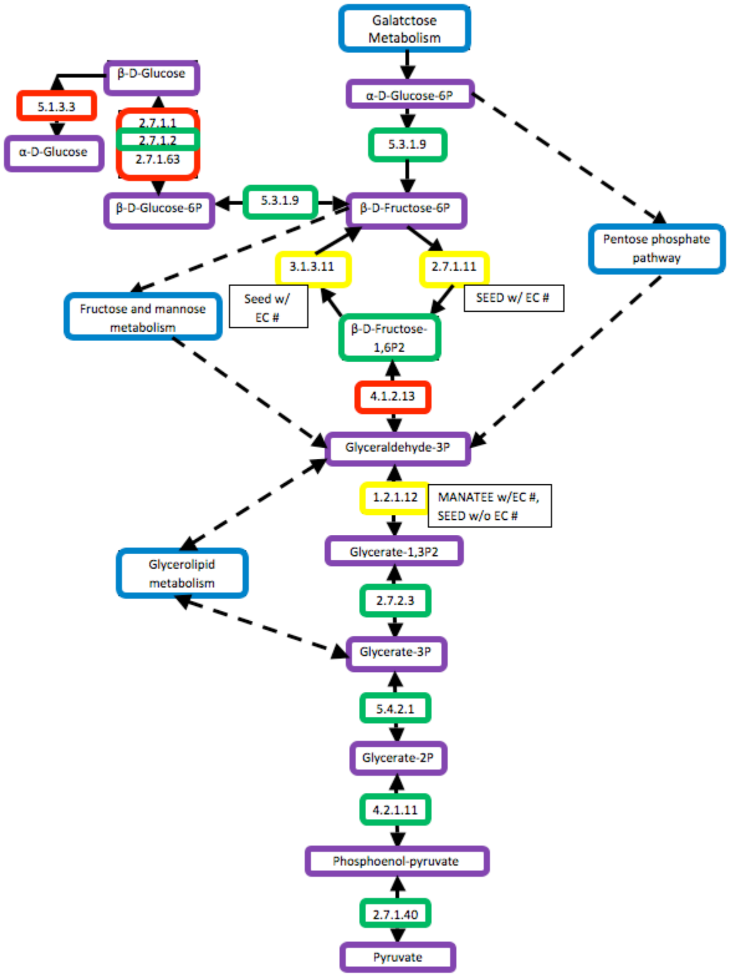
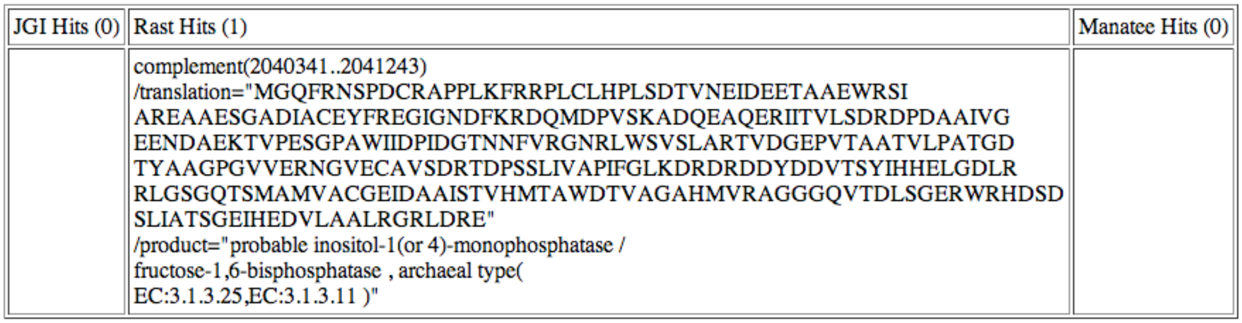


Figure 6. Hand curation of glycolysis/ gluconeogenesis pathway. This pathway is based on the RAST KEGG pathway. It includes the enzymes that are directly involved in producing pyruvate (end product of glycolysis) and glucose (end product of gluconeogenesis). The legend for the figure is as follows: RED (present in no database), GREEN (present in RAST KEGG pathway), YELLOW (no highlighted in RAST KEGG pathway), PURPLE (molecule), BLUE (pathway).

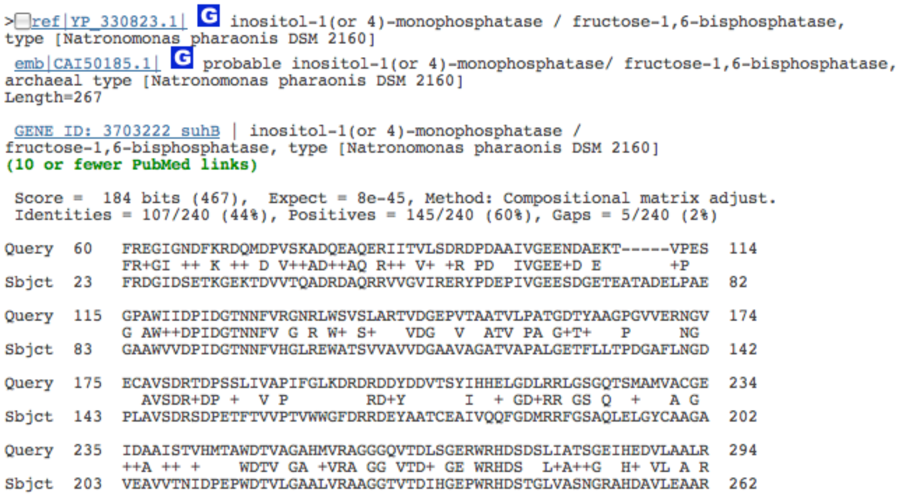
This diagram clearly shows that it is very important to hand curate these pathways. In looking more closely at each gene, I went through a similar procedure as when I was annotating a single ORF. In the above diagram, we can see that enzyme 3.1.3.11 (fructose bisphosphatase) was not highlighted in the RAST KEGG pathway. In order to see whether or not this enzyme was present in any of the other databases, I used the text-search web tool to see if any of the other annotations called this protein.



JGI Hits (0)	Rast Hits (1)	Manatee Hits (0)
	complement(2040341..2041243) /translation="MGQFRNSPDCRAPPLKFRRPLCLHPLSDTVNEIDEETAAEWRSI AREAAESGADIACEYFREGIGNDFKRDQMDPVSKADQEAQERIITVLSDRDPDAAIVG EENDAEKTVPESGPAWIIDPIDGTNNFVRGNRLWSVSLARTVDGEPVTAATVLPATGD TYAAGPGVVERNGVECAVSDRTDPSSLIVAPIFGLKDRDRDDYDDVTSYIHHELGDLR RLGSGQTSMAMVACGEIDAAISTVHMTAWDTVAGAHMVRAGGGQVTDLSGERWRHDSD SLIATSGEIHEDVLAALRGRLDRE" /product="probable inositol-1(or 4)-monophosphatase / fructose-1,6-bisphosphatase , archaeal type(EC:3.1.3.25,EC:3.1.3.11)"	

Figure 7: Results from text search for EC number 3.1.3.11 in the three databases. This diagram shows that the EC number was found in RAST, but not the other two databases. It is, however, slightly ambiguous, because, the database calls two enzymes for the same amino acid sequence.

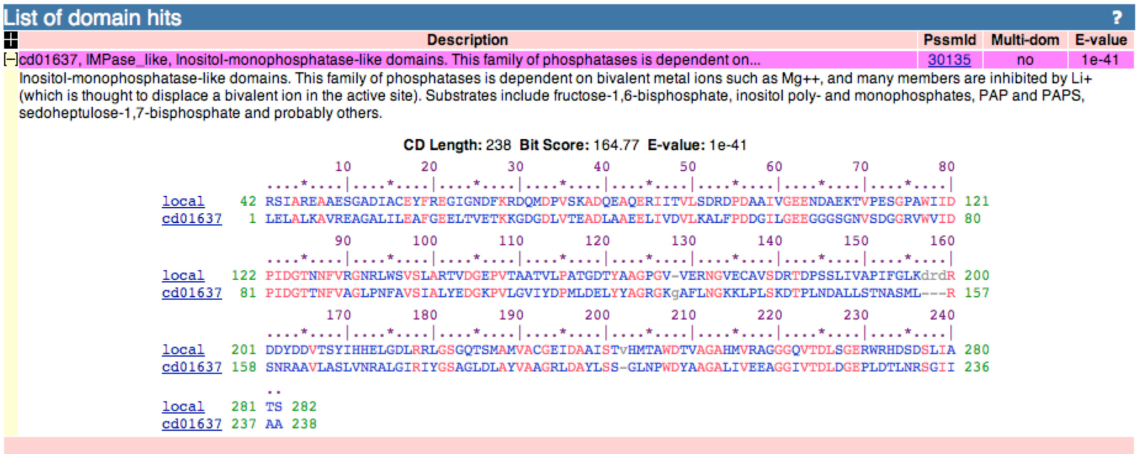
In order to further investigate whether or not this was the correct amino acid sequence for EC number 3.1.3.11, I did a BLASTp search. One of the top alignments came from *Natronomonas pharaonis*, a halobacterium that is closely related to *H. utahensis.* The name assigned to this alignment was the exact same as the name assigned to this protein sequence in RAST, which suggests that the RAST database used BLASTp to annotate this particular gene.



```
>□ref|YP_330823.1| G inositol-1(or 4)-monophosphatase / fructose-1,6-bisphosphatase,
type [Natronomonas pharaonis DSM 2160]
 emb|CAI50185.1| G probable inositol-1(or 4)-monophosphatase/ fructose-1,6-bisphosphatase,
archaeal type [Natronomonas pharaonis DSM 2160]
Length=267

 GENE ID: 3703222 suhB | inositol-1(or 4)-monophosphatase /
fructose-1,6-bisphosphatase, type [Natronomonas pharaonis DSM 2160]
(10 or fewer PubMed links)

 Score =  184 bits (467),  Expect = 8e-45, Method: Compositional matrix adjust.
 Identities = 107/240 (44%), Positives = 145/240 (60%), Gaps = 5/240 (2%)

Query  60   FREGIGNDFKRDQMDPVSKADQEAQERIITVLSDRDPDAAIVGEENDAEKT-----VPES  114
            FR+GI ++ K ++ D V++AD++AQ R++ V+ +R PD  IVGEE+D E       +P
Sbjct  23   FRDGIDSETKGEKTDVVTQADRDAQRRVVGVIRERYPDEPIVGEESDGETEATADELPAE  82

Query  115  GPAWIIDPIDGTNNFVRGNRLWSVSLARTVDGEPVTAATVLPATGDTYAAGPGVVERNGV  174
            G AW++DPIDGTNNFV G R W+ S+   VDG  V  ATV PA G+T+   P     NG
Sbjct  83   GAAWVVDPIDGTNNFVHGLREWATSVVAVVDGAAVAGATVAPALGETFLLTPDGAFLNGD  142

Query  175  ECAVSDRTDPSSLIVAPIFGLKDRDRDDYDDVTSYIHHELGDLRRLGSGQTSMAMVACGE  234
              AVSDR+DP +  V P        RD+Y      I +  GD+RR GS Q +    A G
Sbjct  143  PLAVSDRSDPETFTVVPTVWWGFDRRDEYAATCEAIVQQFGDMRRFGSAQLELGYCAAGA  202

Query  235  IDAAISTVHMTAWDTVAGAHMVRAGGGQVTDLSGERWRHDSDSLIATSGEIHEDVLAALR  294
            ++A ++ +    WDTV GA +VRA GG VTD+ GE WRHDS  L+A++G  H+ VL A R
Sbjct  203  VEAVVTNIDPEPWDTVLGAALVRAAGGTVTDIHGEPWRHDSTGLVASNGRAHDAVLEAAR  262
```

Figure 8. BLASTp alignment for RAST amino acid sequence 3.1.3.11. This alignment has a very low e-value and it comes from an organism that is very closely related to *H. utahensis*. The name of the alignment shows where RAST most likely retrieved the name for this protein call. These results still show some ambiguity in the protein name call.

The BLASTp results still showed some ambiguity in the protein name call. In order to confirm whether or not this enzyme was fructose-1,6-bisphosphatase, I entered the sequence from RAST into CDD and Pfam (result now shown here). Both databases retrieved inositol-monophosphatase domains/superfamilies. The CDD description of this alignment detailed that proteins with this domain have substrates that include fructose-1,6-bisphosphate, which is the molecule that enzyme 3.1.3.11 modifies.



List of domain hits ?

Description	PssmId	Multi-dom	E-value
cd01637, IMPase_like, Inositol-monophosphatase-like domains. This family of phosphatases is dependent on...	30135	no	1e-41

Inositol-monophosphatase-like domains. This family of phosphatases is dependent on bivalent metal ions such as Mg++, and many members are inhibited by Li+ (which is thought to displace a bivalent ion in the active site). Substrates include fructose-1,6-bisphosphate, inositol poly- and monophosphates, PAP and PAPS, sedoheptulose-1,7-bisphosphate and probably others.

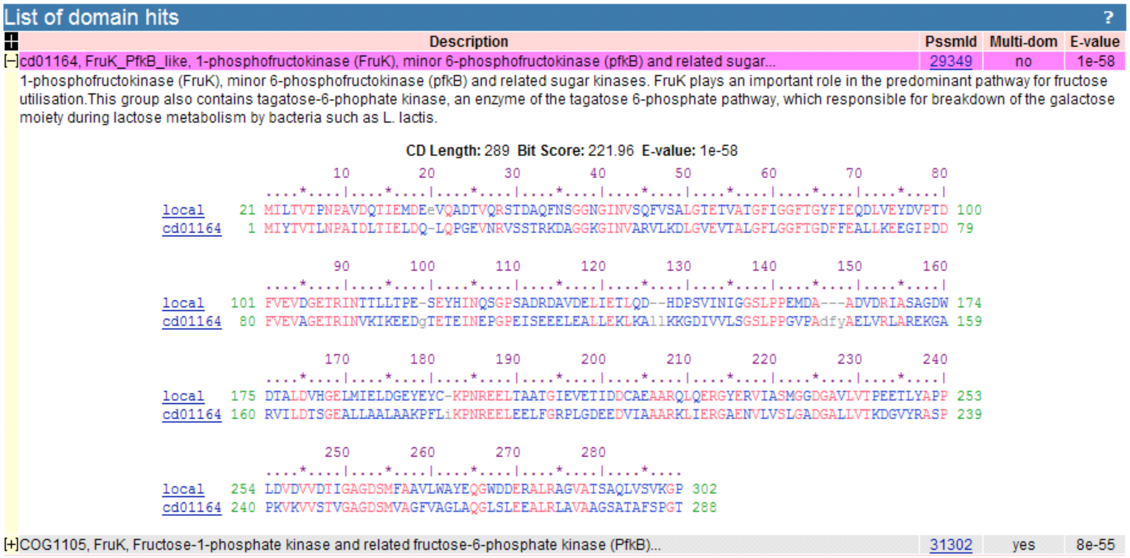
CD Length: 238 **Bit Score:** 164.77 **E-value:** 1e-41

```
                     10        20        30        40        50        60        70        80
             ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local     42 RSIAREAAESGADIACEYFREGIGNDFKRDQMDPVSKADQEAQERIITVLSDRDPDAAIVGEENDAEKTVPESGPAWIID 121
cd01637    1 LELALKAVREAGALILEAFGEELTVETKKGDGDLVTEADLAAEELIVDVLKALFPDDGILGEEGGGSGNVSDGGRVWVID 80
                     90       100       110       120       130       140       150       160
             ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local    122 PIDGTNNFVRGNRLWSVSLARTVDGEPVTAATVLPATGDTYAAGPGV-VERNGVECAVSDRTDPSSLIVAPIFGLKdrdR 200
cd01637   81 PIDGTTNFVAGLPNFAVSIALYEDGKPVLGVIYDPMLDELYYAGRGKgAFLNGKKLPLSKDTPLNDALLSTNASML---R 157
                    170       180       190       200       210       220       230       240
             ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local    201 DDYDDVTSYIHHELGDLRRLGSGQTSMAMVACGEIDAAISTvHMTAWDTVAGAHMVRAGGGQVTDLSGERWRHDSDSLIA 280
cd01637  158 SNRAAVLASLVNRALGIRIYGSAGLDLAYVAAGRLDAYLSS-GLNPWDYAAGALIVEEAGGIVTDLDGEPLDTLNRSGII 236

             ..
local    281 TS 282
cd01637  237 AA 238
```

Figure 9. CDD results for RAST EC number 3.1.3.11. This alignment shows that proteins with this superfamily have substrats that include fructose-1,6-bisphosphatase. In addition, this alignment has a very low e-value. These results suggest that this amino acid sequence is a fructose-1,6-bisphosphatase.

I also looked at the KEGG pathway for *H. salinarum* and found that this enzyme was present in this organism, which supports the presence of EC number 3.1.3.11 in *H. utahensis.* The above process is an example of how to hand curate an enzyme in a pathway to see if the enzymes necessary to complete the pathway are present.

We can see from the hand curated pathway that there are some critical enzymes missing in the glycolysis/gluconeogenesis pathway. I further investigated the enzymes that were not initially highlighted in the RAST KEGG pathway, but that I found in the one of the three databases. One of these enzymes was 2.7.1.11 (6-phosphofructokinase). I located this enzyme in the RAST and Manatee databases using the text-based search tool and found five protein calls between the two databases: two in RAST and three in Manatee. The two sequences in RAST matched with two of the sequences in Manatee. These protein calls were interesting, because, although the calls in the RAST database called EC number 2.7.1.11, the protein call associated with this EC number was 1-phosphofructokinase. Alternatively, Manatee called EC number 2.7.1.56 and protein 1-phosphofructokinase, which is the correct protein name-EC number association. When both of these sequences were entered into CDD, they returned the same result: 1-phosphofructokinase.



List of domain hits ?

Description	PssmId	Multi-dom	E-value
[–]cd01164, FruK_PfkB_like, 1-phosphofructokinase (FruK), minor 6-phosphofructokinase (pfkB) and related sugar...	29349	no	1e-58

1-phosphofructokinase (FruK), minor 6-phosphofructokinase (pfkB) and related sugar kinases. FruK plays an important role in the predominant pathway for fructose utilisation.This group also contains tagatose-6-phophate kinase, an enzyme of the tagatose 6-phosphate pathway, which responsible for breakdown of the galactose moiety during lactose metabolism by bacteria such as L. lactis.

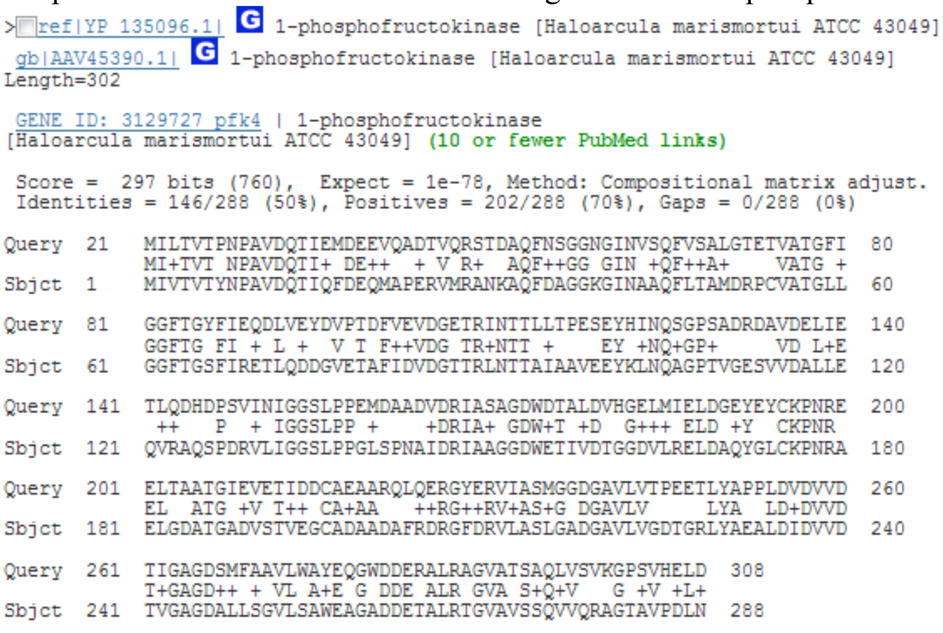
CD Length: 289 **Bit Score:** 221.96 **E-value:** 1e-58

```
                      10        20        30        40        50        60        70        80
                ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local       21  MILTVTPNPAVDQTIEMDEeVQADTVQRSTDAQFNSGGNGINVSQFVSALGTETVATGFIGGFTGYFIEQDLVEYDVPTD 100
cd01164      1  MIYTVTLNPAIDLTIELDQ-LQPGEVNRVSSTRKDAGGKGINVARVLKDLGVEVTALGFLGGFTGDFFEALLKEEGIPDD 79

                      90       100       110       120       130       140       150       160
                ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local      101  FVEVDGETRINTTLLTPE-SEYHINQSGPSADRDAVDELIETLQD--HDPSVINIGGSLPPEMDA---ADVDRIASAGDW 174
cd01164     80  FVEVAGETRINVKIKEEDgTETEINEPGPEISEEELEALLEKLKAllKKGDIVVLSGSLPPGVPAdfyAELVRLAREKGA 159

                     170       180       190       200       210       220       230       240
                ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local      175  DTALDVHGELMIELDGEYEYC-KPNREELTAATGIEVETIDDCAEAARQLQERGYERVIASMGGDGAVLVTPEETLYAPP 253
cd01164    160  RVILDTSGEALLAALAAKPFLiKPNREELEELFGRPLGDEEDVIAAARKLIERGAENVLVSLGADGALLVTKDGVYRASP 239

                     250       260       270       280
                ....*....|....*....|....*....|....*....|....*....
local      254  LDVDVVDTIGAGDSMFAAVLWAYEQGWDDERALRAGVATSAQLVSVKGP 302
cd01164    240  PKVKVVSTVGAGDSMVAGFVAGLAQGLSLEEALRLAVAAGSATAFSPGT 288
```

[+]COG1105, FruK, Fructose-1-phosphate kinase and related fructose-6-phosphate kinase (PfkB)...	31302	yes	8e-55

Figure 10: CDD results for RAST and Manatee phosophofructokinase calls. This alignment shows the results for the amino acid sequences called with EC number 2.7.1.11 and protein name 1-phosphofructokinase in RAST. This alignment shows that the superfamily present in this sequence is associated with 1-phosphofructokinase, which suggests that RAST called the correct protein name, but incorrect EC number.

Next, I entered these sequences into NCBI to try and determine how these databases called these proteins. Both of the sequences returned very similar results and the top alignment hit for both sequences came from *Haloarcula marismortui*, a halobacterium that is closely related to *H. utahensis*. The protein name calls for both of these alignments was 1-phosphofructokinase.



```
>□ref|YP_135096.1| G 1-phosphofructokinase [Haloarcula marismortui ATCC 43049]
 gb|AAV45390.1| G 1-phosphofructokinase [Haloarcula marismortui ATCC 43049]
Length=302

 GENE ID: 3129727 pfk4 | 1-phosphofructokinase
[Haloarcula marismortui ATCC 43049] (10 or fewer PubMed links)

 Score =  297 bits (760),  Expect = 1e-78, Method: Compositional matrix adjust.
 Identities = 146/288 (50%), Positives = 202/288 (70%), Gaps = 0/288 (0%)

Query  21   MILTVTPNPAVDQTIEMDEEVQADTVQRSTDAQFNSGGNGINVSQFVSALGTETVATGFI  80
            MI+TVT NPAVDQTI+ DE++  + V R+  AQF++GG GIN +QF++A+    VATG +
Sbjct  1    MIVTVTYNPAVDQTIQFDEQMAPERVMRANKAQFDAGGKGINAAQFLTAMDRPCVATGLL  60

Query  81   GGFTGYFIEQDLVEYDVPTDFVEVDGETRINTTLLTPESEYHINQSGPSADRDAVDELIE  140
            GGFTG FI + L +  V T F++VDG TR+NTT +    EY +NQ+GP+     VD L+E
Sbjct  61   GGFTGSFIRETLQDDGVETAFIDVDGTTRLNTTAIAAVEEYKLNQAGPTVGESVVDALLE  120

Query  141  TLQDHDPSVINIGGSLPPEMDAADVDRIASAGDWDTALDVHGELMIELDGEYEYCKPNRE  200
             ++   P  + IGGSLPP +    +DRIA+ GDW+T +D  G+++ ELD +Y  CKPNR
Sbjct  121  QVRAQSPDRVLIGGSLPPGLSPNAIDRIAAGGDWETIVDTGGDVLRELDAQYGLCKPNRA  180

Query  201  ELTAATGIEVETIDDCAEAARQLQERGYERVIASMGGDGAVLVTPEETLYAPPLDVDVVD  260
            EL  ATG +V T++ CA+AA   ++RG++RV+AS+G DGAVLV     LYA  LD+DVVD
Sbjct  181  ELGDATGADVSTVEGCADAADAFRDRGFDRVLASLGADGAVLVGDTGRLYAEALDIDVVD  240

Query  261  TIGAGDSMFAAVLWAYEQGWDDERALRAGVATSAQLVSVKGPSVHELD  308
            T+GAGD++ + VL A+E G DDE ALR GVA S+Q+V   G +V +L+
Sbjct  241  TVGAGDALLSGVLSAWEAGADDETALRTGVAVSSQVVQRAGTAVPDLN  288
```

Figure 11: NCBI BLASTp results for sequence matches in RAST and Manatee for phosphofructokinase. This is the top alignment when both of the sequences that matched in RAST and Manatee were entered into BLASTp. The low e-value and relatively high identity percentage show that these amino acid sequences are most likely 1-phosphofructokinases.

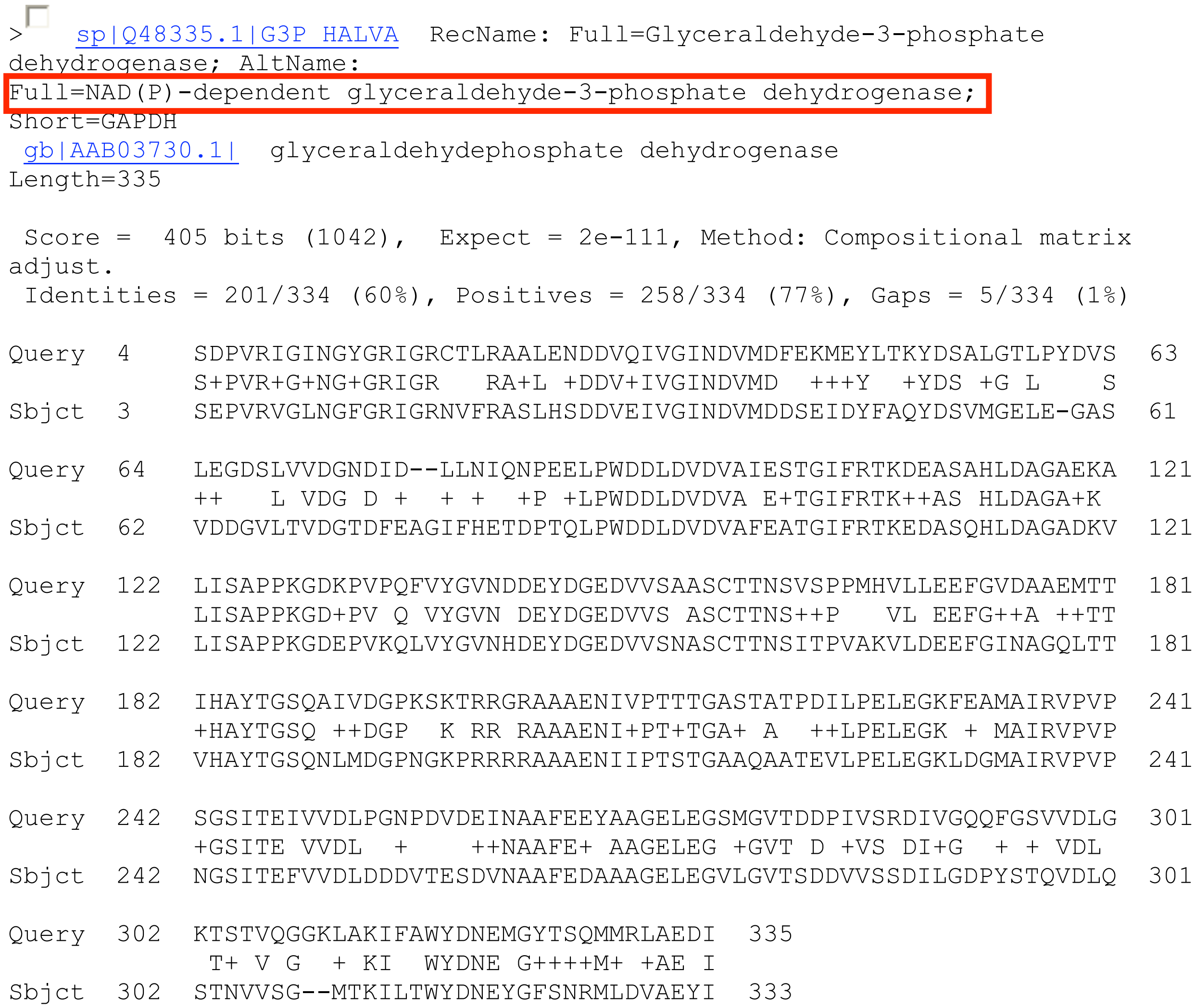
In an attempt to further understand the discrepancy in EC numbers calls for the two amino acid sequences that matched in RAST and Manatee, I compared these sequences with amino acid sequences associated with EC numbers 2.7.1.11 and 2.7.1.56. For the amino acid sequence associated with EC number 2.7.1.11, I had searched in KEGG pathway for an organism that his enzyme and used that amino acid sequence. Even though the sequence for enzyme 2.7.1.11 came from a different organism, it should still have conserved domains that would make it a 6-phosphofructokinase. When I compared the first hit that matched between the RAST and Manatee databases to the enzyme 2.7.1.11, using BLAST2, the results returned no significant similarity. The same was true for when I compared enzyme 2.7.1.11 to the second hit that matched in the RAST and Manatee databases. Because the BLAST2 alignments did not return any significant similarity, this suggests that the amino acid sequences in RAST, which were called with EC number 2.7.1.11, do not have the functional domains that would make them a 6-phosphofructokinase.

Next I compared the sequences that matched between RAST and Manatee to the amino acid sequence of EC number 2.7.1.56 (1-phosphofructokinase), which I retrieved from the organism *Vibrio fischeri*. Both of these comparisons returned alignments that had very low e-values and seemed to have domains of amino acid conservation. This suggests that the amino acid sequences in RAST that called EC number 2.7.1.11, are most likely associated with EC number 2.7.1.56.

```
Score =  127 bits (318),  Expect = 3e-27
 Identities = 96/301 (31%), Positives = 145/301 (48%), Gaps = 18/301 (5%)

Query  22   ILTVTPNPAVDQTIEMDEEVQADTVQRSTDAQFNSGGNGINVSQFVSALGTETVATGFIG  81
            ++T+T NPA+D T  +D  +   +V   +    ++ G G+NV++ +S LG E   TGF+G
Sbjct  9    VVTITLNPALDLTGSLDA-LSVGSVSLVSKGSLHAAGKGVNVAKVLSDLGAEVTVTGFLG  67

Query  82   GFTGYFIEQDLVEYDVPTDFVEVDGETRINTTLLTPESEYH-INQSGPSADRDAVDE---  137
                    Q   E      F+ VDG TRIN  L+  +     IN  G    + A+ E
Sbjct  68   RDNEELFCQLFEEMKAKDQFIRVDGATRINVKLVESDGRVSDINFPGVEVSQQAIAEFEV  127

Query  138  -LIETLQDHDPSVINIGGSLP----PEMDAADVDRIASAGDWDTALDVHGELMIELDGEY  192
             L E  +DHD  V+   GSLP    PE  A  ++++   G           L   LD
Sbjct  128  RLFELAKDHDFFVL--AGSLPKGISPEQCAEWIEKLHQMGKKVLFDSSRAALAAGLDAHP  185

Query  193  EYCKPNREELTAATGIEVETIDDCAEAARQLQERGYERVIASMGGDGAVLV-----TPEE  247
               KPN EEL+   G E+ T + C +AA  L E+G E ++ S+G  G + +        E
Sbjct  186  WLIKPNDEELSEFVGRELNTPESCQQAAEDLAEKGIENIVVSLGSKGVMWLGQNNQEQAE  245

Query  248  TLYAPPLDVDVVDTIGAGDSMFAAVLWAY-EQGWDDERALRAGVATSAQLVSVKGPSVHE  306
             +Y+ P  ++VV T+GAGD++ A + W + +Q WD  + L    A SA  VS  G  V +
Sbjct  246  WMYSQPPKMNVVSTVGAGDTLVAGLCWGHMQQDWDRSQILSFATALSALAVSQVGVGVPD  305

Query  307  L  307
            +
Sbjct  306  I  306

Score =  119 bits (299),  Expect = 4e-25
 Identities = 90/301 (29%), Positives = 139/301 (46%), Gaps = 18/301 (5%)

Query  2    IVTVTLNPAVDQTIKMNTGLQSGSVQRSTEAQFTSGGNGVNVSQFLQALGSETVATGLIG  61
            +VT+TLNPA+D T  ++  L  GSV   ++    + G GVNV++ L  LG+E   TG +G
Sbjct  9    VVTITLNPALDLTGSLDA-LSVGSVSLVSKGSLHAAGKGVNVAKVLSDLGAEVTVTGFLG  67

Query  62   GFTGYFIENDLATYDVSTDFVWVEGVTRINTTILTPRNEYQ-LNQTGPTVDSDVIDE---  117
                               F+ V+G TRIN  ++        +N  G  V    I E
Sbjct  68   RDNEELFCQLFEEMKAKDQFIRVDGATRINVKLVESDGRVSDINFPGVEVSQQAIAEFEV  127

Query  118  -LIEIISQHDPDTLNIGGSLLPGMD----AADVDRIATAGDWDTAVEVPGEVLSELDADY  172
             L E+   HD   L   GSL  G+     A  ++++   G           + + LDA
Sbjct  128  RLFELAKDHDFFVL--AGSLPKGISPEQCAEWIEKLHQMGKKVLFDSSRAALAAGLDAHP  185

Query  173  AYCKPNREELEAATGHEIDSVTDCVDAAKTLQERGFECVIASMGSEGAVMV-----TPEE  227
               KPN EEL    G E+++   C  AA+ L E+G E ++ S+GS+G + +        E
Sbjct  186  WLIKPNDEELSEFVGRELNTPESCQQAAEDLAEKGIENIVVSLGSKGVMWLGQNNQEQAE  245

Query  228  TLYAPALDVEVVDTLGAGDSMLAAVLWAR-EQGWDAERALRAGVVASAQLVGVMGSSVRE  286
             +Y+    + VV T+GAGD+++A + W   +Q WD  + L      SA  V  +G  V +
Sbjct  246  WMYSQPPKMNVVSTVGAGDTLVAGLCWGHMQQDWDRSQILSFATALSALAVSQVGVGVPD  305

Query  287  L  287
            +
Sbjct  306  I  306
```

Figure 12: Alignment between phosphofructokinase hit matches in SEED and Manatee and enzyme 2.7.1.56 in *V. fischeri.* This alignment shows that the amino acid sequences that were called as 2.7.1.11 (6-phosphofructokinase) in the RAST database, actually seem to have conserved domains that make them associated with EC number 2.7.1.56 (1-phosphofructokinase). Both alignments also give very low e-values.

I followed a similar procedure to investigate enzyme 1.2.1.12, which is a glyceraldehyde-3-phosphate dehydrogenase. This enzyme was present in all three databases and all three databases had the exact same amino acid sequence. However, RAST called the EC number 1.2.1.59, which is also a glyceraldehyde-3-phoshpate dehydrogenase, but this enzyme uses the cofactor NADP$^+$. I first entered the amino acid sequence into BLASTp. The first hit was an NADP-dependent glyceraldehyde-3-phosphate dehydrogenase with a very low e-value.



```
  ☐
>   sp|Q48335.1|G3P_HALVA  RecName: Full=Glyceraldehyde-3-phosphate
dehydrogenase; AltName:
Full=NAD(P)-dependent glyceraldehyde-3-phosphate dehydrogenase;
Short=GAPDH
 gb|AAB03730.1|  glyceraldehydephosphate dehydrogenase
Length=335

 Score =  405 bits (1042),  Expect = 2e-111, Method: Compositional matrix
adjust.
 Identities = 201/334 (60%), Positives = 258/334 (77%), Gaps = 5/334 (1%)

Query  4    SDPVRIGINGYGRIGRCTLRAALENDDVQIVGINDVMDFEKMEYLTKYDSALGTLPYDVS  63
            S+PVR+G+NG+GRIGR   RA+L +DDV+IVGINDVMD  +++Y  +YDS +G L    S
Sbjct  3    SEPVRVGLNGFGRIGRNVFRASLHSDDVEIVGINDVMDDSEIDYFAQYDSVMGELE-GAS  61

Query  64   LEGDSLVVDGNDID--LLNIQNPEELPWDDLDVDVAIESTGIFRTKDEASAHLDAGAEKA  121
            ++   L VDG D +  + +  +P +LPWDDLDVDVA E+TGIFRTK++AS HLDAGA+K
Sbjct  62   VDDGVLTVDGTDFEAGIFHETDPTQLPWDDLDVDVAFEATGIFRTKEDASQHLDAGADKV  121

Query  122  LISAPPKGDKPVPQFVYGVNDDEYDGEDVVSAASCTTNSVSPPMHVLLEEFGVDAAEMTT  181
            LISAPPKGD+PV Q VYGVN DEYDGEDVVS ASCTTNS++P   VL EEFG++A ++TT
Sbjct  122  LISAPPKGDEPVKQLVYGVNHDEYDGEDVVSNASCTTNSITPVAKVLDEEFGINAGQLTT  181

Query  182  IHAYTGSQAIVDGPKSKTRRGRAAAENIVPTTTGASTATPDILPELEGKFEAMAIRVPVP  241
            +HAYTGSQ ++DGP  K RR RAAAENI+PT+TGA+ A  ++LPELEGK + MAIRVPVP
Sbjct  182  VHAYTGSQNLMDGPNGKPRRRRAAAENIIPTSTGAAQAATEVLPELEGKLDGMAIRVPVP  241

Query  242  SGSITEIVVDLPGNPDVDEINAAFEEYAAGELEGSMGVTDDPIVSRDIVGQQFGSVVDLG  301
            +GSITE VVDL  +    ++NAAFE+ AAGELEG +GVT D +VS DI+G  + + VDL
Sbjct  242  NGSITEFVVDLDDDVTESDVNAAFEDAAAGELEGVLGVTSDDVVSSDILGDPYSTQVDLQ  301

Query  302  KTSTVQGGKLAKIFAWYDNEMGYTSQMMRLAEDI  335
             T+ V G  + KI  WYDNE G++++M+ +AE I
Sbjct  302  STNVVSG--MTKILTWYDNEYGFSNRMLDVAEYI  333
```

Figure 13: BLASTp alignment for glyceraldehyde-3-phosphate dehydrogenase amino acid sequence found in databases. This alignment is the first hit in the search and it retrieves a very low e-value (2e-111). The red box shows a potential reason for why RAST may have associated this amino acid sequence with enzyme 1.2.1.59.

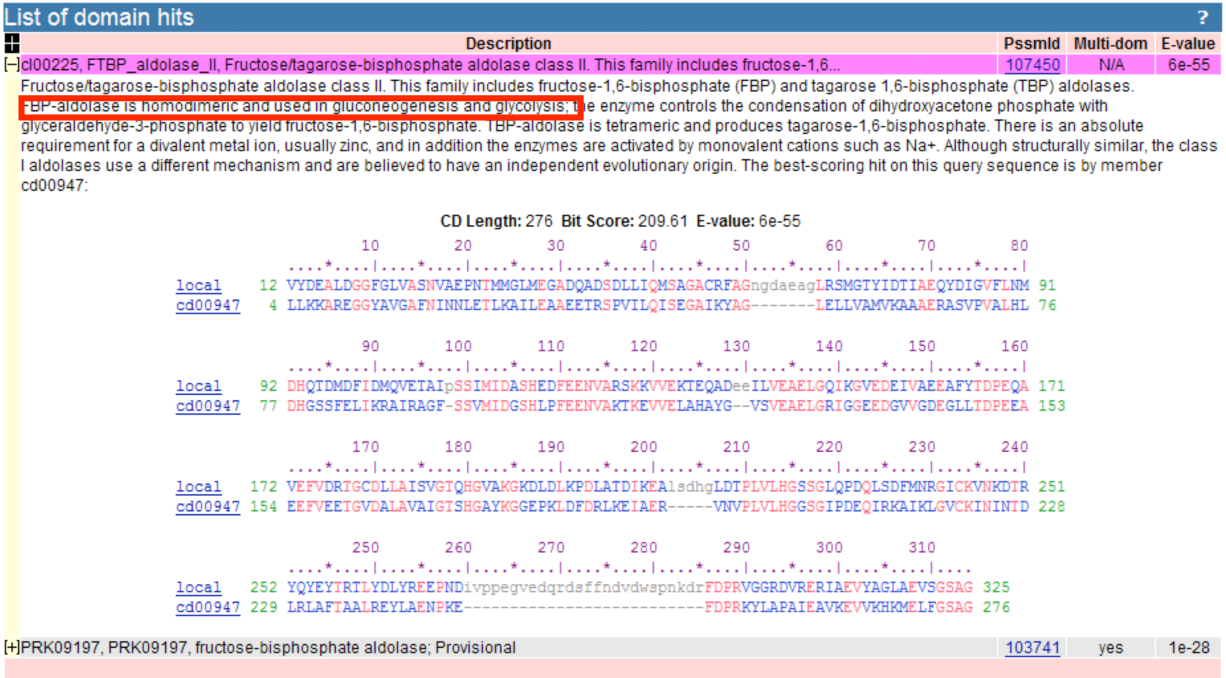
In order to further investigate the findings from this alignment, I compared the amino acid sequence found in the three databases to the amino acid sequences associated with enzymes 1.2.1.12 and 1.2.1.59. For enzyme 1.2.1.59 (for which I could not find in a pathway in the KEGG pathway database) I used an amino acid sequence associated with enzyme 1.2.1.13 because these are both glyceraldehyde-3-phosphate dehydrogenase (NADP$^+$ dependent). Both alignments gave very low e-values. However, it is interesting that the amino acid sequence associated with enzyme 1.2.1.13 produced such a statistically significant alignment, because this enzyme in generally found in photosynthetic organisms. It is not, however, surprising that both of these alignments returned very similar alignments, because the enzymes 1.2.1.12 and 1.2.1.59 have very similar functions, and thus have very similar conserved domains. In addition, Manatee called the EC number as 1.2.1-, meaning that it could not determine which type of glyceraldehyde-3-phosphate this amino acid sequence was. This could be due to the fact that the amino acid sequence in the databases is closely related to both enzyme 1.2.1.12 and enzyme 1.2.1.59. For this reason, I believe that this is a case where we will not be able to definitely determine which EC number this amino acid sequence is related unless we perform experiments for the presence of both of these enzymes in our organism.

```
Score =  257 bits (657),  Expect = 1e-66
 Identities = 130/307 (42%), Positives = 191/307 (62%), Gaps = 4/307 (1%)

Query  28   NDDVQIVGINDVMDFEKMEYLTKYDSALGTLPYDVSLEGDSLVVDGNDIDLLNIQNPEEL  87
            N D+ IV IND+ D + + +L KYDS    LP ++    +S+++DG +I +   ++PE L
Sbjct  23   NSDIDIVAINDLTDAKTLAHLFKYDSVHKILPNEIKATENSIIIDGKEIKIFAEKDPENL  82

Query  88   PWDDLDVDVAIESTGIFRTKDEASAHLDAGAEKALISAPPKGDKPVPQFVYGVNDDEYDG  147
            PW DL++DV +ESTG+FR ++ A  HL AGA+K +I+AP KG+      V G N+++
Sbjct  83   PWKDLNIDVVVESTGVFRNREGAEKHLKAGAKKVVITAPAKGEDIT--IVLGCNEEQLKP  140

Query  148  ED-VVSAASCTTNSVSPPMHVLLEEFGVDAAEMTTIHAYTGSQAIVDGPKSKTRRGRAAA  206
            E  ++S ASCTTNS++    V+ +EF +    + T+H+YT  Q I+D P    RR RAAA
Sbjct  141  EHKIISCASCTTNSIASIAKVINDEFKIITGHLITVHSYTNDQRILDLPHKDLRRARAAA  200

Query  207  ENIVPTTTGASTATPDILPELEGKFEAMAIRVPVPSGSITEIVVDLPGNPDVDEINAAFE  266
             NI+PTTTGA+ A   ++PEL+GK + MAIRVP P GS+T + V +      +E+N   +
Sbjct  201  VNIIPTTTGAAKAVALVVPELKGKLDGMAIRVPTPDGSLTNLSVIVEKATTAEEVNEVVK  260

Query  267  EYAAGELEGSMGVTDDPIVSRDIVGQQFGSVVDLGKTSTVQGGKLAKIFAWYDNEMGYTS  326
            +   G L+G +G   +PIVS DIVG  +  + D   T  V  G L  IF+WYDNE GYT
Sbjct  261  KATEGRLKGIIGYNTEPIVSGDIVGTTYAGIFDATLTK-VMNGNLVNIFSWYDNEYGYTC  319

Query  327  QMMRLAE  333
            +++   E
Sbjct  320  RVVDTLE  326

Score =  236 bits (601),  Expect = 5e-60
 Identities = 142/331 (42%), Positives = 208/331 (62%), Gaps = 7/331 (2%)

Query  21   LKVAINGFGRIGRNFLRCWHGRKDSPLDVIVVNDTGGVKQASHLLKYDSILGTFEADVKA  80
            +++ ING+GRIGR  LR     ++  + ++ +ND    ++  +L KYDS LGT   DV
Sbjct  7    VRIGINGYGRIGRCTLRA--ALENDDVQIVGINDVMDFEKMEYLTKYDSALGTLPYDVSL  64

Query  81   VGDDAISVDGKVIKIVSSRNPLDLPWGDLDIDLVIEGTGVFVDRDGAGKHIQAGAKKVLI  140
             GD  + VDG  I +++ +NP +LPW DLD+D+ IE TG+F  +D A  H+ AGA+K LI
Sbjct  65   EGDSLV-VDGNDIDLLNIQNPEELPWDDLDVDVAIESTGIFRTKDEASAHLDAGAEKALI  123

Query  141  TAPGKGD--IPTYVVGVNADEYNHDESIISNASCTTNCLAPFVKVLDQKFGIIKGTMTTT  198
            +AP KGD  +P +V GVN DEY+  E ++S ASCTTN ++P + VL ++FG+    MTT
Sbjct  124  SAPPKGDKPVPQFVYGVNDDEYD-GEDVVSAASCTTNSVSPPMHVLLEEFGVDAAEMTTI  182

Query  199  HSYTGDQRLLDASHRDLRRARAAALNIVPTSTGAAKAVALVLPTLKGKLNGIALRVPTPN  258
            H+YTG Q ++D       RR RAAA NIVPT+TGA+ A   +LP L+GK   +A+RVP P+
Sbjct  183  HAYTGSQAIVDGPKSKTRRGRAAAENIVPTTTGASTATPDILPELEGKFEAMAIRVPVPS  242

Query  259  VSVVDLVVQVSKKTFAEEVNAGFRDSAEKELQGILSVCDEPLVSVDF-RCSDVSSTVDSS  317
             S+ ++VV +      +E+NA F + A  EL+G + V D+P+VS D       S
Sbjct  243  GSITEIVVDLPGNPDVDEINAAFEEYAAGELEGSMGVTDDPIVSRDIVGQQFGSVVDLGK  302

Query  318  LTMVMGDDMVKVIAWYDNEWGYSQRVVDLAD  348
             + V G  + K+ AWYDNE GY+ +++ LA+
Sbjct  303  TSTVQGGKLAKIFAWYDNEMGYTSQMMRLAE  333
```

Figure 14: Alignments between glyceraldehyde-3-phosphate amino acid sequences and EC number 1.2.1.12 and 1.2.1.59. The top alignment is a comparison between the amino acid sequence found in the databases and the amino acid sequence for enzyme 1.2.1.12 (the amino acid sequence came from

Thermosipho melanesiensis). The bottom alignment is a comparison between the amino acid sequence found in the databases and the amino acid sequence associated with enzyme 1.2.1.13 (which has the same name associated with it as enzyme 1.2.1.59-the amino acid sequence came from *Vitis vinifera*). Both of these alignments give very low e-values, which makes it difficult to determine which enzyme is more likely to be present in our organism.

With this initial comparison of the glycolysis/gluconeogenesis pathway, it seems that enzyme 4.1.2.13 (fructose bisphosphate aldolase) and enzyme 5.1.3.3 (aldose 1-epimerase) do not exist in our organism (colored red in Figure 6). However, in an attempt to make sure that the protein calls made by the database were accurate, we also searched for these enzymes using its alternative names in the three databases. For enzyme 4.1.2.13 I found two hits, one in RAST and one in Manatee for fructose/tagatose-1,6-bisphosphate aldolase when performing a text search for "aldolase" and no hits when searching for the other alternative names found in the ExPASy search engine. Both of the protein sequences found in RAST and Manatee were identical. In order to confirm that this protein was in fact enzyme 4.1.2.13, I entered the amino acid sequence into CDD. The domain hit in CDD was a fructose/tagatose-bisphosphate aldolase.


List of domain hits ?

Description	PssmId	Multi-dom	E-value
[-]cl00225, FTBP_aldolase_II, Fructose/tagarose-bisphosphate aldolase class II. This family includes fructose-1,6...	107450	N/A	6e-55

Fructose/tagarose-bisphosphate aldolase class II. This family includes fructose-1,6-bisphosphate (FBP) and tagarose 1,6-bisphosphate (TBP) aldolases. FBP-aldolase is homodimeric and used in gluconeogenesis and glycolysis; the enzyme controls the condensation of dihydroxyacetone phosphate with glyceraldehyde-3-phosphate to yield fructose-1,6-bisphosphate. TBP-aldolase is tetrameric and produces tagarose-1,6-bisphosphate. There is an absolute requirement for a divalent metal ion, usually zinc, and in addition the enzymes are activated by monovalent cations such as Na+. Although structurally similar, the class I aldolases use a different mechanism and are believed to have an independent evolutionary origin. The best-scoring hit on this query sequence is by member cd00947:

CD Length: 276 **Bit Score:** 209.61 **E-value:** 6e-55

```
                    10        20        30        40        50        60        70        80
           ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local     12 VYDEALDGGFGLVASNVAEPNTMMGLMEGADQADSDLLIQMSAGACRFAGngdaeagLRSMGTYIDTIAEQYDIGVFLNM 91
cd00947    4 LLKKAREGGYAVGAFNINNLETLKAILEAAEETRSPVILQISEGAIKYAG-------LELLVAMVKAAAERASVPVALHL 76

                    90       100       110       120       130       140       150       160
           ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local     92 DHQTDMDFIDMQVETAIpSSIMIDASHEDFEENVARSKKVVEKTEQADeeILVEAELGQIKGVEDEIVAEEAFYTDPEQA 171
cd00947   77 DHGSSFELIKRAIRAGF-SSVMIDGSHLPFEENVAKTKEVVELAHAYG--VSVEAELGRIGGEEDGVVGDEGLLTDPEEA 153

                   170       180       190       200       210       220       230       240
           ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
local    172 VEFVDRTGCDLLAISVGTQHGVAKGKDLDLKPDLATDIKEAlsdhgLDTPLVLHGSSGLQPDQLSDFMNRGICKVNKDTR 251
cd00947  154 EEFVEETGVDALAVAIGTSHGAYKGGEPKLDFDRLKEIAER-----VNVPLVLHGGSGIPDEQIRKAIKLGVCKININTD 228

                   250       260       270       280       290       300       310
           ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....
local    252 YQYEYTRTLYDLYREEPNDivppegvedqrdsffndvdwspnkdrFDPRVGGRDVRERIAEVYAGLAEVSGSAG 325
cd00947  229 LRLAFTAALREYLAENPKE--------------------------FDPRKYLAPAIEAVKEVVKHKMELFGSAG 276
```

[+]PRK09197, PRK09197, fructose-bisphosphate aldolase; Provisional	103741	yes	1e-28

Figure 15. Fructose bisphosphate aldolase protein sequence in CDD. This figure shows that the superfamily associated with this sequence contains fructose bisphosphate aldolase. The description highlighted in red confirms that this protein is used in both glycolysis and gluconeogenesis.

I also did a search for alternative names for enzyme 5.1.3.3 (aldose 1-epimerase) and found no hits in any of the databases. Even though this organism does not have enzyme 5.1.3.3 does not seem to be a hindrance, because this enzyme's job is to convert β-glucose to α-glucose and vice versa. Both forms of glucose provide energy they just have slightly different molecular structures.

Discussion:

The investigation of individual genes in the *H. utahensis* genome provides a great deal of information, both about the organism itself and about the databases' annotation abilities. One of the results from this investigation showed that, even though an annotation database may name a protein, this does not necessarily mean that the name is correlated to the function. We saw this in the example of the peroxiredoxin vs. the monooxygenase, in which RAST called this ORF a monooxygenase and JGI and Manatee called it a peroxiredoxin. It seems that the RAST database called this ORF a monooxygenase, because this enzyme was the first hit in the BLASTp results. It does not seem that this database checks to validate that these calls are consistent with the functional domains present in the sequence itself because the monooxygenase call in the RAST database retrieved a conserved domain that was related to peroxiredoxins. This result strengthens the idea that using multiple annotation databases and other online tools to validate the calls made by an individual database is very important.

Another important lesson from this investigation was the fact that the RAST KEGG pathway does not accurately portray the enzymes that are present in certain metabolic pathways in our organism. It seems that this database only highlights EC numbers if the EC number is part of the title of an ORF and does not highlight the EC number even when the protein name corresponding to the EC number in the pathway is present in the annotation. This necessitates manual investigation to determine whether or not the EC number is present in the annotation and this requires a great deal of work on the investigator's part. In manually curating the glycolysis/gluconeogenesis pathway, I found that, in some cases, the databases called a different EC number with the same amino acid sequence. When I looked at 6-phosphofructokinase in the glycolysis/gluconeogenesis pathway, RAST called EC number 2.7.1.11 and Manatee called EC number 2.7.1.56 for the same amino acid sequences. This may have occurred because both of these proteins have very similar functions, but it is very ambiguous for the databases to associate different EC numbers with the same amino acid sequence. In this case, it led to the conclusion that the enzyme 6-phosphofructokinase may not be present in *H. utahensis* at this locus. In this example, I also found that the databases sometimes associate a protein name with the wrong EC number. For example, EC number 2.7.1.11, which is 6-phosphofructokinase, was associated with a protein sequence that was named 1-phosphopfructokinase. In further investigating this sequence, I found that it is most likely a 1-phosphofructokinase. This discrepancy may have also occurred because these two proteins are very closely related. However, if the goal of annotating a genome is to gain an accurate picture of a genome, these discrepancies are significant. This shows that solely relying on one annotation database could prove to be very problematic and it also illustrates the idea that the three annotation tools use different methods to assign EC numbers and protein names.

In investigating the glyceraldehyde-3-phosphate dehydrogenase in the glycolysis/gluconeogenesis pathway, I found that the databases do not always completely name an EC number if there is some ambiguity as to which EC number the amino acid sequence is associated with. This can occur in cases like what I saw with EC numbers1.2.1.12 (glyceraldehyde-3-phosphate dehydrogenase) and 1.2.1.59 (glyceraldehyde-3-phosphate dehydrogenase NADP$^+$ dependentManatee only called the EC number as 1.2.1-, while RAST called 1.2.1.59 and JGI called 1.2.1.12. In my investigation of this discrepancy, I found that the alignments between each of the amino acid sequences associated with these EC numbers and the amino acid sequence found in the three databases are very close (they both have very low EC numbers), which may be why Manatee did not definitively call a specific glyceraldehdye dehydrogenase. This again shows that certain annotation tools have different methods of assigning EC numbers and that some may be more cautious in assigning numbers when there are protein sequences that are so closely related. However, it also shows that these databases need to be altered in order to be able to provide complete and accurate information about an ORF.

Another important consideration in manually investigating the protein calls made by annotation databases is to look at alternative names for a specific protein. EC number 4.1.2.13 (fructose bisphosphate aldolase) did not appear in the results for the text-based search tool, until we searched the databases using an alternative name. Not being able to search for a protein without knowing all of its alternative names can be a great hindrance in trying to annotate a novel genome. In the future, it would be beneficial for the annotation database to compile all the alternative names for a protein when it names a specific ORF.

In comparing the annotation abilities of the three databases when looking at specific genes and pathways, it does not seem that any of the databases annotate the genome better than the others. All of them make mistakes in calling proteins and all of them have benefits that the others do not. In general, I believe that JGI provides the best information if you are looking for general information about the organism's genome and if you want to manually search for alternative ORFs. This database is also very useful for finding other information related to a specific gene, like the COGs or protein families associated with that gene. RAST is very useful in looking at the metabolic pathways in the organism. It is also very useful if you are interested in looking at the subsystems that enzymes belong to. Manatee is very useful in searching for a specific gene. It is however, rather difficult to navigate this website, because you have to have a very good idea of what you are looking for beforehand.

The exploration of the *H. utahensis* genome has not only led to a greater understanding of a unique organism, but has also revealed the intricacies of using online annotation tools. Overall, we found that it is very difficult to assume that the protein calls made by these databases are final. In all of the specific examples I looked at, there were multiple instances when the databases did not agree with each other. For this reason, we deem that it is important to conduct further online investigation to confirm the calls the databases make.

In the future, it would be beneficial to complete the comparison of the three databases for this organism in order to more completely understand how the databases differ in their genome annotation. This information may reveal patterns in how certain annotation engines call proteins, which could be beneficial in creating a tool that would pull information from a database based on what the database does well. It would also be beneficial to perform experiments to confirm whether or not certain enzymes are present in *H. utahensis*. These experiments could include testing growth in different kinds of media or directly testing for the presence of certain enzymes. This information would give a much more complete picture of both the genome of *Halorhabdus utahensis* and the annotation abilities of JGI, RAST and Manatee.

Acknowledgements:
We would like to thank our contacts at JGI, Cheryl Kerfeld and Edwin Kim. We would like to thank Dr. Jonathan Eisen at UC Davis and Dr. Gary Stormo at Washington University for guidance during this project. We would like to thank Matt DeJongh at Hope College for RAST for assisting us in using the RAST server and Ramana Madupu at the J. Craig Venter Institute for doing the same with Manatee database. We would especially like to thank Kjeld Ingvorsen at the Biologisk Insitut at Aarhus Universitet in Denmark for collaborating with us and conducting experiments with *H. utahensis.* Finally, we would like to thank Chris Healey at Davidson College for ordering and growing *H. utahensis* locally.

References:

1. Alqueres SMC, Almeida RV, Clementino MM, Vieira RP, Almeida WI, Cardoso AM, Martins OB (2007) Exploring the biotechnological applications in the archaeal domain. Brazilian Journal of Microbiology 38: 398-405.

2. Hingamp P, Brochier C, Talla E, Gautheret D, Thieffry D, Herrmann (2008). Metagenome annotation using a distributed grid of undergraduate students. PLoS Biology 6: 2362-2367.

3. Schiraldi C, Guiliano M, De Rosa M (2002) Perspectives on biotechnological applications of Archaea. Archaea 1:75-86.

4. Waino M, Tindall BJ, Ingvorsen K (2000) Halorhabdus utanhesis gen. nov., sp nov., an Aerobic extremely halophilic member of the Archaea from Great Salt Lake, Utah. International Journal of Systematic and Evolutionary Microbiology 50: 183-190.