**Analyzing the genome of *Halorhabdus utahensis* using three different databases**

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

Department of Biology, Davidson College
[1]Department of Mathematics, Davidson College

**Abstract**

One way to harvest information about a species is to automatically annotate its genome using a specialized computer tool, and then curate the genome manually based on the tool's annotation. Several annotation websites exist that accept sequenced genomes, but they are not all the same. The *Halorhabdus utahensis* genome was submitted to three tools to be automatically annotated: RAST, TIGR's Manatee server, and JGI's Integrated Microbial Genomes' (IMG) site. The three servers predicted different genes due to internal biases and theories governing the creation of the annotation tool. RAST, for example, is more likely to predict longer genes that use alternative start codons (not ATG). The differences in annotation tools highlight genes of interest – genes predicted by only one or two of the databases – that require further study. After annotating *H. utahensis*, pathway analyses allow the user to infer function based on genes that are predicted in the genome.

**Introduction**

Archaea are single-celled prokaryotes that make up a domain in the three-domain system. They are separated from the bacteria due to their unusual biochemistry and appearance – outwardly, they resemble prokaryotes, but they have metabolic pathways similar to those of eukaryotes. Halophiles are a class of archaea known for their ability to tolerate high

salt concentrations. *Halorhabdus utahensis*, isolated from Great Salt Lake in Utah, is an aerobic halophile that optimally grows in 27% (w/v) NaCl (1). Comparing the analysis of this genome on three databases – The Doe Joint Genome Institute (JGI), Manatee, and Rapid Annotation using Subsystem Technology (RAST) – highlights discrepancies between the databases and unique aspects of *Halorhabdus utahensis*'s genome. Comparing the databases' similar findings and looking for outliers can streamline identification and classification of all genes and pathways. We are conducting this research in an undergraduate class, and we have found it helpful to divide the responsibilities of genome annotation amongst us (2). The divide and conquer method lead to a mastery of tools and a greater, more efficient understanding of the *H. utahensis* genome.

**Materials and Methods**

To find sequence similarities, NCBI's Blastx and Blastn were used.

To identify dissimilarities amongst called proteins in the three databases, the results of a pairwise comparison of genes on the largest of five DNA scaffolds (3102403 bps of 3129561 total bps) were analyzed. The largest scaffold had a GC content of 62.9%. The other scaffolds had GC contents of 62.2%, 58.2%, 63.4%, and 65.5%. The pairwise comparison can be found here: http://gcat.davidson.edu/Registry/compare/.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) was used to analyze pathways and compare the *H. utahensis* pathways to other known halophiles.

RAST provides an annotated gene list and website for analysis via the SEED Viewer. Tools available included a breakdown of genes called by subsystem, a BLAST search of the *H. utahensis* genome, KEGG metabolic analysis of the genome, and a genome browser.

JGI'S IMG provides an annotated gene list and a website for analysis. Tools include a table breaking down the called ORFs into different types of genes, a genome browser and chromosome maps, easy links to other databases on the gene page, and the ability to export genome information in FASTA format.

Manatee provides separate annotated gene lists for analysis on a searchable website.

**Results**

*Database Comparison*

JGI called 3097 open reading frames (ORFs), RAST called 2898 ORFs, and Manatee called 3254 ORFs on the largest DNA scaffold. To attempt to identify the discrepancy between the number of ORFs according to each annotation website, a BioPerl computer program was written to compare the websites' called start and stop sites. The ORFs that matched exactly are illustrated in figure 1 below.

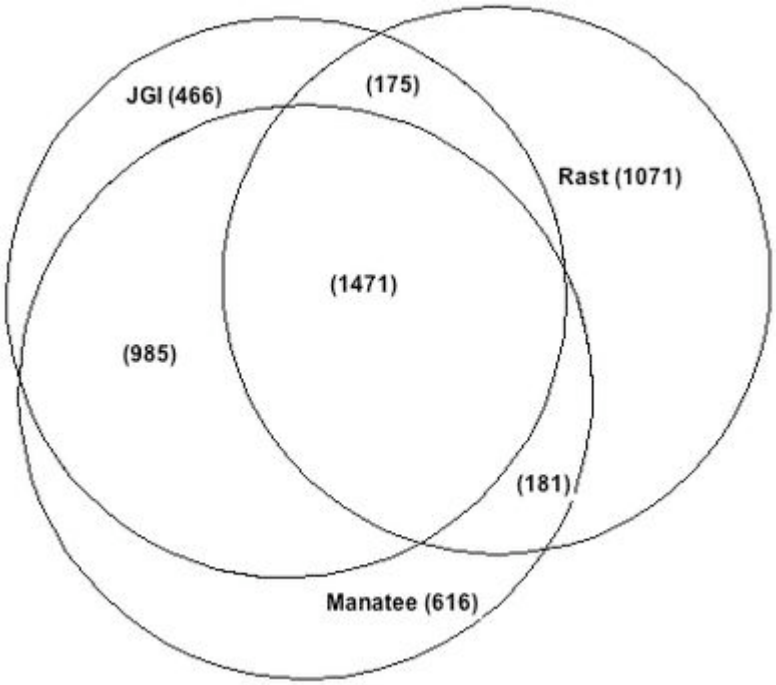



**Figure 1**. Venn diagram illustrating exact start and stop DNA matches determined by BioPerl comparison algorithm. JGI and Manatee matched each other 2456 times, which is significantly more than RAST ORFs match to either JGI or Manatee ORFs.

To seek an explanation as to why there was so much discrepancy between the three databases, we decided to do a less rigorous comparison using only stop codon matches. Start codons can be either ATG, TTG, GTG, or CTG and some variance can exist in the calling of start codons, while the first stop codon invariably signals the end of a protein transcript. A second Venn diagram showing the stop codon matches was generated (Figure 2).

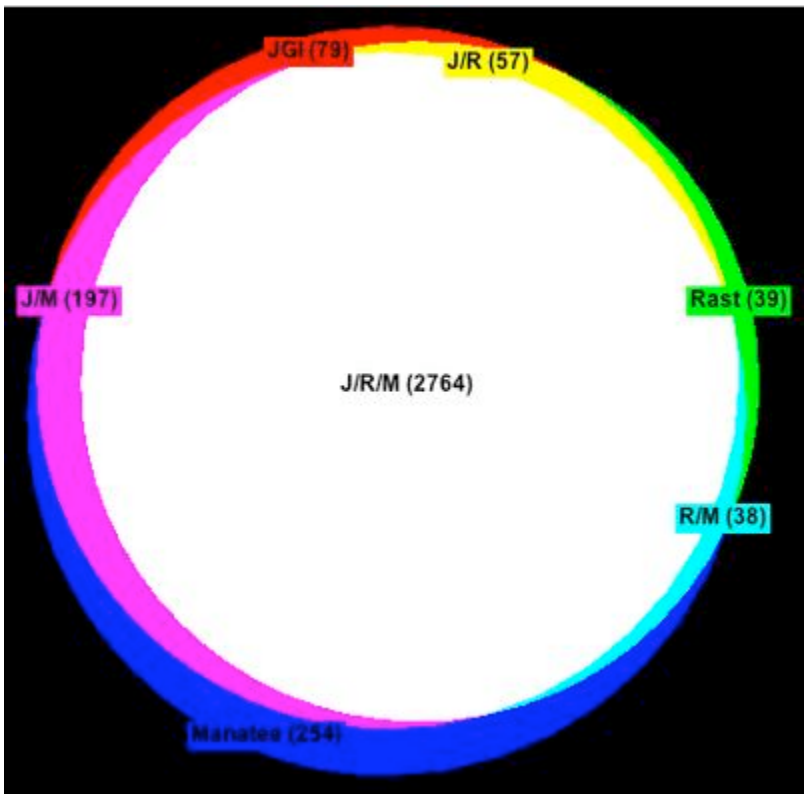



**Figure 2**. Venn diagram showing stop codon matches between JGI (J), Manatee (M), and RAST (R) ORF annotations. The number of matches between JGI and Manatee is no longer significantly more than RAST ORF matches to either JGI or Manatee.

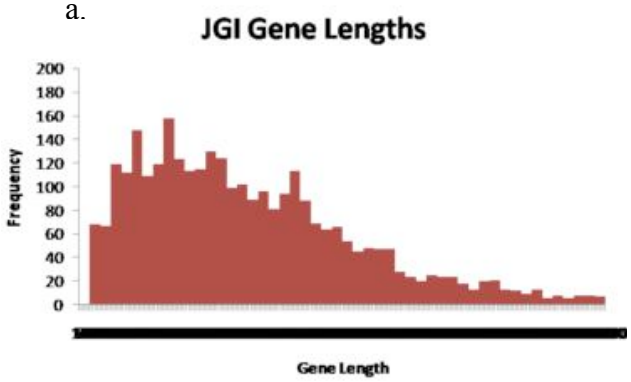
When only stop codons are considered, 2764 ORFs match in all three comparisons, yet when both stop and start codons are considered, 1071 ORFs are found only by RAST. To determine why this occurred, a BioPerl program was written to calculate the number of each type of start codon (ATG, GTG, TTG, CTG, or another) identified by all three annotation websites (Table 1).

| 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% |

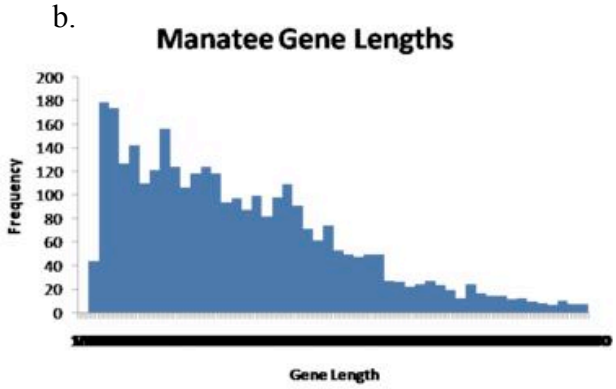
**Table 1**. Results of start codon analysis. ATG is the most commonly used start codon in all three annotations, however, 40.5% of ORFs in RAST started with an alternative start codon. This may partially explain why RAST had 1071 ORFs that did not match both start and stop codons called by JGI and Manatee, but only 39 ORFs that did not match the stop codons. Note: No ORFs starting with CTG were found – 94% of alternative start codons used GTG or TTG. RNA genes are included.

A BioPerl program was written to determine predicted ORF lengths as determined by all three annotation websites (Figure 3). Differing gene length trends may help explain why RAST called so many unique genes when both stop and start codons were considered.

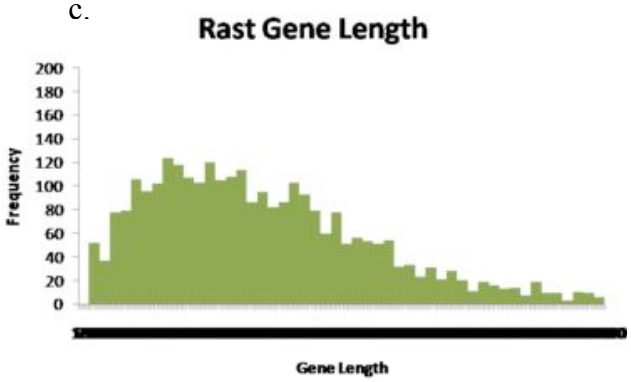
a.



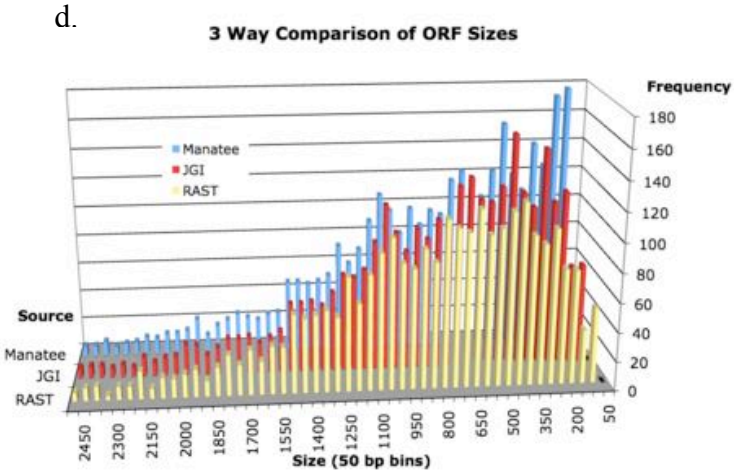

b.




c.




d.




e.

| Statistic | JGI | RAST | Manatee |
|---|---|---|---|
| Mean | 869.9 | 941.8 | 844.9 |
| Median | 728 | 801.5 | 692 |
| Mode | 428 | 284 | 116 |
| Minimum | 70 | 70 | 73 |
| Maximum | 7130 | 10001 | 10001 |
| Count | 3097 | 2898 | 3254 |

**Figure 3**. Distribution and comparison of predicted ORF lengths of all annotations. X-axis for graphs a, b, and c range from 0-2500 base pairs with a 50-base pair step value. Table e numerically displays the results of the ORF length analysis. RAST predicted ORF lengths are significantly longer than JGI and Manatee predicted ORFs, which may explain why RAST had 1071 ORFs that did not match both start and stop codons called by JGI and Manatee, but only 39 ORFs that did not match the stop codons.

Overall, trends indicated that RAST called longer genes. We decided to break "genes" into two groups: genes where all three databases called the same stop codons, and genes that were unique to a database (Figure 4).

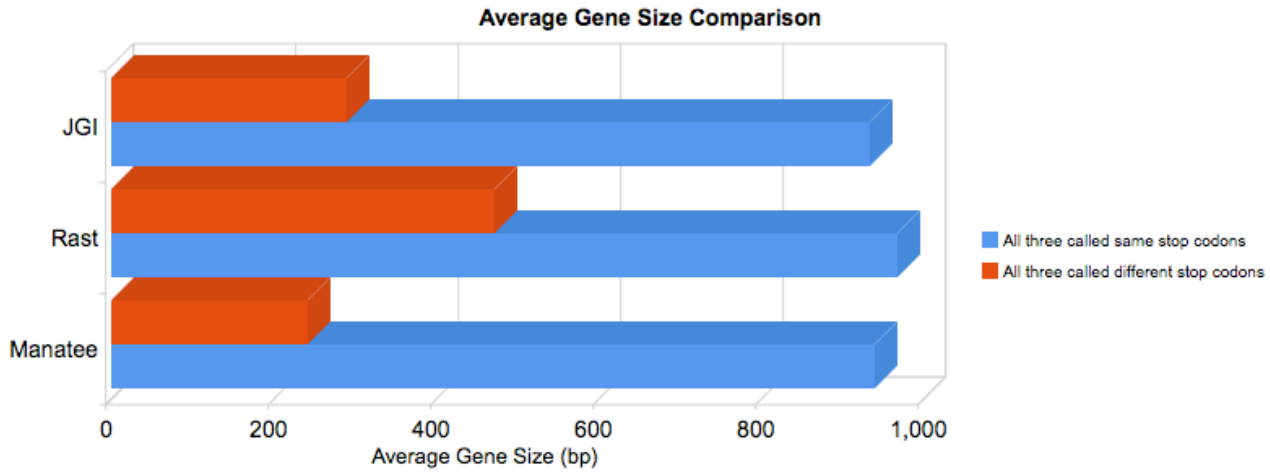



**Figure 4.** Gene size comparisons of genes that shared the same stop codons, and of genes that were unique to a database. When considering genes with the same stop codons, JGI had an average length of 934, RAST 967, and Manatee 940, all of which are similar. However, when considering unique genes, RAST's average length was 472, almost two times as long as Manatee's average of 242 and JGI's average of 290.

This data shows that the three databases seem to be in agreement about ORF length when comparing genes that they all found. Genes that were unique to different databases dramatically vary in length, yet are significantly shorter than the genes shared by all three databases. The different in ORF length may indicate that they are less reliable gene calls than the shared gene calls.

*Single Gene Analysis*

Discrepancies between JGI, Manatee, and RAST ORFs required closer analysis so several unique ORFs were examined singularly. For example, JGI gene 2500587699, a 375-bp hypothetical protein from base pairs 80504-80878 on the positive strand, did not share its start or stop codon sites with any ORFs called by Manatee or RAST. An NCBI blastx search with the nucleotide sequence showed that the first half of the sequence

matched FtsZ2 from several halobacterium, and the second half of the sequence matched CopG from several halobacterium (Figure 5).


**Figure 5**. Blastx alignment scores for bps 80504-80878 in the H. utahensis genome. The first column are matches with FtsZ genes, and the second column are alignment matches with CopG genes.

After refining the *H. utahensis* nucleotide sequences searched, base pairs 79962-81193 on the negative strand matched the FtsZ2 nucleotide sequence from *Haloarcula japonica* with an e-value of 0.0. However, single base pair additions and deletions starting at the 354$^{th}$ base pair in the sequence may result in a null protein. JGI predicts 2 copies of FtsZ, a necessary protein required for cell division, one of which ranges from base pairs 79500-80261 on the negative strand (less than 300 base pairs away from the hypothetical protein JGI called on the positive strand) (3). JGI may want to elongate the length of the FtsZ protein it calls to include some of the bases from the hypothetical protein it identified, as these appear to be part of FtsZ. Manatee and RAST call FtsZ2 from base pairs 79500-80102 on the negative strand, and FtsZ from base pairs 80102-80701 on the negative strand, which may be more accurate than JGI's prediction. Base pairs 80708-80896 matched the CopG sequence from *Halorubrum lacusprofundi* with an e-value of 4x10^-17 when performing a blastx. JGI predicts 7 CopG family proteins. Manatee predicts 5 CopG family proteins and 1 protein as actually being CopG, a protein that binds to dilysine motifs and assists with protein transport. No predicted CopG family proteins overlap the predicted region from 80708-80896, which may indicate an additional protein not predicted by any of three annotation databases. It is unclear why JGI did not correctly predict the FtsZ proteins.

*Database Literature Review*

To further identify differences in the annotation databases, we looked at publications and presentations produced by JGI, Manatee, and RAST. JGI's tool is called the integrated microbial genomes (IMG) system. Highlights of this tool include its user-

friendliness and integration with other available genomes that allows for easy comparison (4). Every gene in IMG is characterized by COG membership, Pfam domain, Gene Ontology assignment, and KEGG enzyme associations. All of these annotations can be searched by keyword, which contributes to JGI's user-friendliness. IMG's genome annotations are fully integrated with NCBI Entrez Gene. After identifying a predicted ORF, IMG performs an NCBI BLASTp to find the most similar homolog match, as long as the e-value is less than 0.01 (5).

RAST was created to allow experts to annotate subsystems using a decentralized approach (6). The central idea to RAST annotation is that experts on particular subsystems will help annotate that subsystem in various species. The subsystems are determined by FIGfams, and a particular organism's FIGfams are viewed in a subsystem spreadsheet (7). If a gene is missing from the spreadsheet, it can be manually curated by an expert on that subsystem. RAST also allows its expert users to have easy access to information needed to determine the accuracy of its bioinformatics programs (8). RAST calls genes by finding all ORFs and then identifying the start codon by aligning the ORF with known genes (6).

Manatee relies on sequence similarity to existing known genes when calling proteins, and will typically manually curate genomes after the automatic curation. Manatee annotation is efficient, using BLAST to identify similar proteins to the ORF in question, and only performing the Smith-Waterman algorithm on proteins with a certain match value (9). Manatee also includes a Genome Property report page that predicts the presence or absence of pathways and structures of a particular genome.

*Pathway Analyses: Amylases and Purines*

The *H. utahensis* genome hopefully gives the user insight into what the organism should be able to do. Analysis of pathways is possible with RAST's KEGG metabolic analysis viewer, but the tool is not as accurate as it could be. For example, looking at the Starch and Sucrose metabolism pathway in *H. utahensis*, all three enzymes used to degrade starch are not present in the KEGG map. However, the RAST protein spreadsheet identifies two of the three enzymes as present in the annotations.

*H. utahensis* appears to have an almost complete purine metabolism pathway (Figure 5).

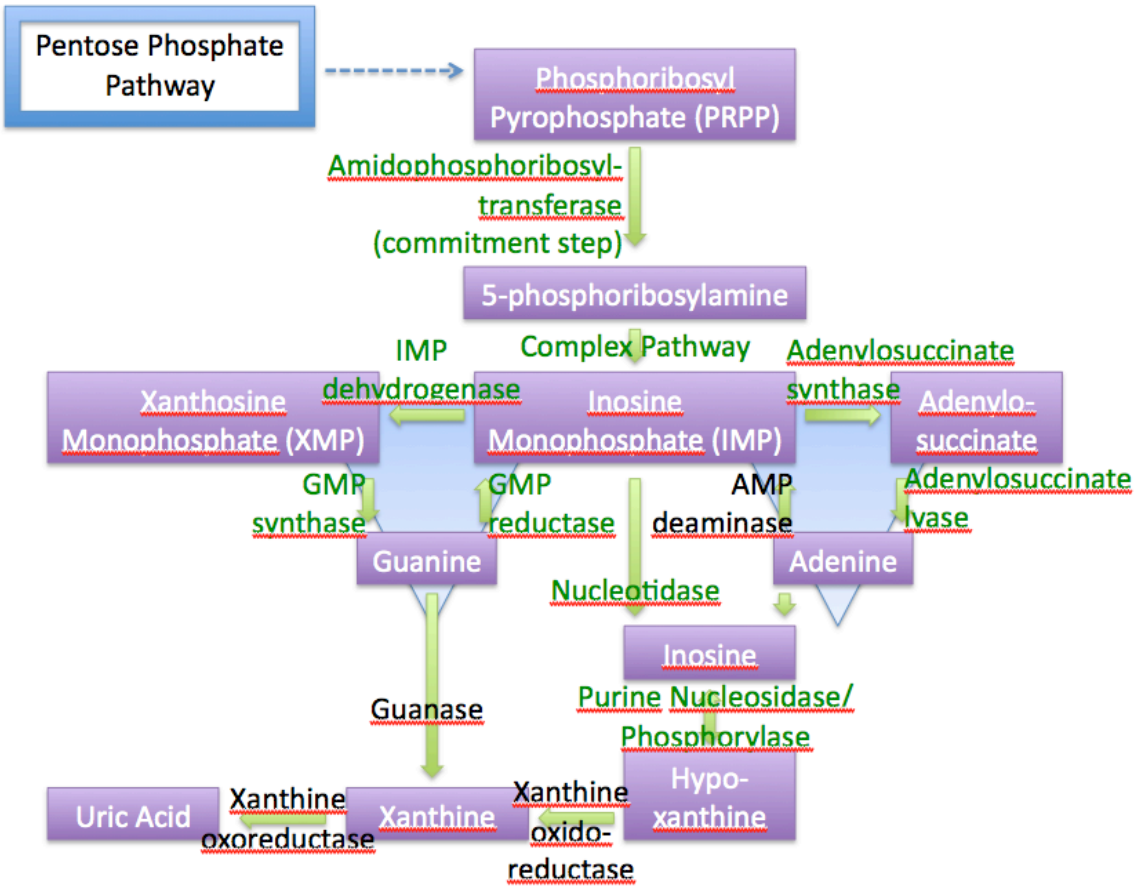



**Figure 5**. The purine metabolism pathway in *H. utahensis*. The Pentose Phosphate Pathway (blue box) feeds into this pathway. Purple boxes are molecules. Green arrows represent enzymes. Green writing means *H. utahensis* has this enzyme, while black writing means it appears not to have that enzyme.

It is missing xanthine reductases, which turn xanthine into urea, but these are not necessary in archaea as xanthine is the final waste product of the cell and urea does not need to be produced. *H. utahensis* is also missing an AMP deaminase and a guanase, unless it has a protein with a specialized, previously unknown function that performs as an AMP deaminase or a guanase. The AMP deaminase is only needed to convert adenine back to IMP, which only occurs when adenine is not converted to inosine and then to xanthine and then removed from the cell. Guanase is only needed to produce xanthine, which is not made in *H. utahensis*. Thus, just by annotating the *H. utahensis* genome, one can predict its ability to use starch as a carbon source and its ability to metabolize and excrete purines.

**Discussion**

We further classified *H. utahensis*'s genome and analyzed three commonly used annotation databases to critically assess their differences and similarities. We used the largest DNA scaffold (99.1% of the total genome) reported by sequencing for our analysis, but since the GC contents of all five scaffolds were similar (ranging from 58.2-65.5%), we do not believe this created a bias. In our species, Manatee called the most ORFs and RAST called the least. RAST was more likely to call ORFs with alternative start codons, and RAST called the longest ORFs, especially when looking at the genes RAST uniquely called. This may be due to an internal bias that sets cut-off values for sequence similarity too high, or overlooks smaller genes. It is interesting to note that RAST calls the least ORFs, but its ORFs are the longest, while Manatee calls the most, yet shortest, ORFs. All three sites predicted that roughly 87% of the genome was coding.

Further analysis should be done to optimize the process of calling ORFs at each of the three websites to avoid calling ORFs that are not really genes, but still calling all the genes present. In addition to varying the genes called, the databases also varied in accessibility and information presented. Manatee was the most difficult database to master. JGI's IMG database is easily accessible with many resources available. RAST is also accessible and has a few different resources than the IMG database – we found the KEGG analysis especially helpful, although we found mistakes, such as proteins being present in the RAST annotations but not on the KEGG RAST enzyme map.

By examining the annotation results, many predictions could be made about *H. utahensis*'s function. *H. utahensis* should be able to survive when grown on starches, and should be able to produce its own amino acids and metabolize purines. These predictions can be tested by growing *H. utahensis* on a media whose sole carbon source is starch, on a media devoid of amino acids, and testing the media surrounding *H. utahensis* for xanthine, the purine metabolism waste product. By utilizing all three genome annotation sites, we were able to produce a clear genomic picture of several metabolic pathways which will inform lab-driven experiments. If lab experiments support our model, we have demonstrated the effectiveness of inferring metabolic abilities from enzymes present in the genome. If the lab experiments do not support our predictions, we must revise the model and critically determine where we erred. Using the annotation databases to inform our predictions underlines the importance of having strong computer models on which to base biological predictions. In this case, the strength of the computer models comes in numbers. Having three separate databases annotate our genome provided us with a wealth of information about our genome, but also about the steps needed to create an even

stronger, integrated annotation system. Since the three annotation databases gave us different results, this points to a need for more research in the area of automatic genome annotation. Hopefully the process will become more streamlined to save biologists even more time in predicting and testing how an organism functions.

Being able to predict how an organism functions is important in areas of energy and health research. An organism may be able to utilize a previously unknown molecule as a source of energy, or a new enzyme may be discovered that catalyzes a reaction that advances medicine. Analyzing new genomes with a streamlined annotation database will lead to further insights.

**References**


1. Wainø M, Tindall BJ, and Ingvorsen K (2000) *Halorhabdus utahensis* gen. nov., sp. nov., an aerobic, extremely halophilic member of the *Archaea* from Great Salt Lake, Utah. *International Journal of Systemic and Evolutionary Microbiology* 50:183-190.
2. Hingamp, P. (2008) Metagenome annotation using a distributed grid of undergraduate students. PLOS Biology, 6.11:2362-67
3. Haydon DJ, et al. (2008) An inhibitor of FtsZ with potent and selective anti-staphylococcal activity. *Science*, 321:1673-75.
4. Markowitz, VM et al. (2005) The integrated microbial genome (IMG) system. *Nucleic Acids Research*, 34:4-8.
5. Markowitz VM, et al. (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Research*, 36.5:28-33.

6. Overbeek R, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33.17.

7. Aziz RK, et al. (2008) The RAST server: Rapid annotations using subsystems technology. BMC Genomics, 9.75.

8. DeJongh M, et al. (2007) Toward automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics*, 8.139.

9. Giglio, MG (2005) Prokaryotic Annotation at TIGR. Available < http://www.geneontology.org/teaching_resources/presentations/2005-06_AnnotCamp_TIGR_MGwinn.ppt>.

**Acknowledgments:**

We would like to thank our contacts at JGI, Cheryl Kerfeld and Edwin Kim; our contact at the SEED and RAST, Matt DeJongh of Hope College; and our contact at the J. Craig Venter Institute for Manatee, Ramana Madupu.

We would like to thank Kjeld Ingvorsen, of Det Naturvidenskabelige Fakultet, Biologisk Institut, Aarhus Universitet, Denmark for collaboration in testing our hypotheses on *H. utahensis*

We would like to thank Jonathan Eisen at UC Davis and Gary Stormo of Washington University for their input and guidance in this project.

Finally, we extend our thanks to Chris Healey at Davidson College for ordering and growing *H. utahensis* locally.