A Three-way Comparison of JGI, RAST, and Manatee Annotation Engines by Analysis of the *Halorhabdus utahensis* Genome

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

Abstract

We ran the *Halorhabdus utahensis* genome through three different annotation engines—IMG, RAST, and Manatee—to analyze the similarities and differences between the resulting annotations. We found differences between the annotations to be significant, encompassing fields such as EC number agreement, gene length prediction, start codon prediction, and average gene size, which in turn have an effect on many aspects of the annotation, such as predicted ribosomal binding sites. We also performed numerous case studies, of which two are described in this paper, of specific genes and pathways to pinpoint ways in which the annotations differ or are lacking in functionality or effectiveness.

Introduction

Strain AX-2 of *Halorhabdus utahensis* was isolated from Great Salt Lake in Utah. It is a halophile and a newly described species and genus of the domain *Archaea*; it is related to the family *Halobacteriaceae*, although not similar enough to warrant inclusion in that family.

We wished to examine and compare the annotations produced by three different annotation engines: the Joint Genome Institute's IMG, the SEED's RAST, and J. Craig Venter Institute's Manatee. To get a framework of comparison with which to describe the sorts of calls each engine made, we had each service annotate the genome of *Halorhabdus utahensis*; since the exact same genome sequence was provided to each service, we could highlight the similarities and the differences.

Materials and Methods

The genome of *H. utahensis* was sequenced by the Joint Genome Institute (JGI); they provided us with the scaffolds and nucleotide sequences. We submitted the DNA sequence to three different annotation engines. The first was JGI's in-house annotation engine, Integrated Microbial Genomes (IMG) (http://imgweb.jgi-psf.org/cgi-bin/img_edu_v260/main.cgi). The second was Rapid Annotations using Subsystem Technology (RAST) from the SEED (http://rast.nmpdr.org/seedviewer.cgi?page=Home).

The third was the annotation service of the J. Craig Venter Institute (JCVI), where it was run through JCVI's prokaryotic annotation pipeline. We manually reviewed the output using Manatee (http://www.tigr.org/tigr-scripts/prok_manatee/shared/login.cgi) and the prokaryotic pipeline of the JCVI Annotation Service.

The SEED database possesses one especially useful tool for studying metabolic pathways: the KEGG pathway map. This will show the KEGG pathways for any genome in the SEED database, highlighting each gene that has been called and showing all of the genes that were not called. This provides the user with a simple tool to see which pathways are complete, which are incomplete, etc. We looked at factors such as gene lengths, incorrect pathway analysis, etc. to reveal annotation inconsistencies between the engines.

We used many other tools to aid in our investigation, both public and private. Besides the three annotation engines, other public tools we used included **Pfam** (http://pfam.sanger.ac.uk/), **CDD** (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), **PDB** (http://www.rcsb.org/pdb/home/home.do), **BLAST** (http://blast.ncbi.nlm.nih.gov/Blast.cgi), **KEGG** (http://www.genome.jp/kegg/), **ENZYME** (http://www.expasy.ch/enzyme/), **ExPASy** (http://www.expasy.ch/), and **PubMed** (http://www.ncbi.nlm.nih.gov/pubmed/).

We also wrote several in-house programs, mostly using the BioPerl programming language. These included a tool allowing the user to search for any particular EC number in the annotations of JGI, RAST, or Manatee (http://www.bio.davidson.edu/courses/genomics/2008/Win/ec/), which greatly simplified the task of researching searching for EC number calls across databases; a tool allowing the user to perform a text-based search for protein calls in all three databases (http://gcat.davidson.edu/Wideloache/Webfiles/AnnotationSearcher.html), which simplified text-based searches; and a tool allowing the user to blast an EC number against the *H. utahensis* genome (http://gcat.davidson.edu/Wideloache/Webfiles/ecNumBlast.html), returning any genes that matched below a certain e-value one of the known genes with that EC number. These tools were developed to increase the efficiency of our analyses of the similarities and inconsistencies between the three annotation engines; they greatly aided us in pinpointing specific gene calls and enzymes that were similar or inconsistent between annotations. All tools are available publicly at http://gcat.davidson.edu/GcatWiki/index.php/Halorhabdus_utahensis_Genome , though they are all specific to our genome.

Results

We compared the gene calls made by the three annotation engines. The three annotation engines agreed on the positions of both the start and stop codons for only about half of the genes (Figure 1). The engines generally agreed, however, on the stop codons calls for the genes (Figure 2); the disagreements were largely on where the start codon was.

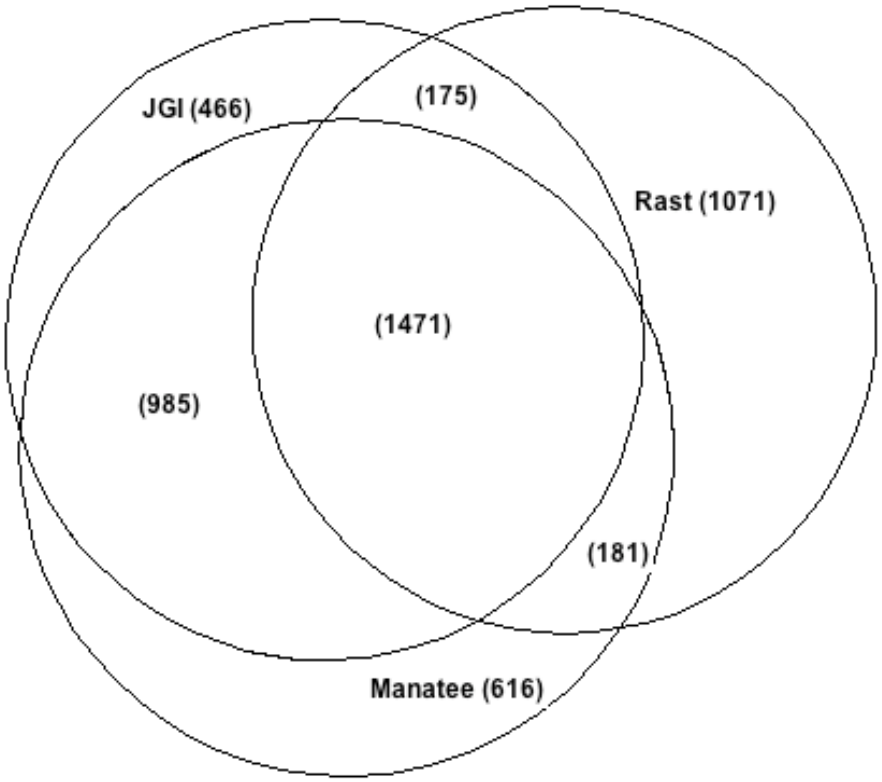

Figure 1 - Venn diagram showing the number of exact gene matches across the three annotations. Regions that overlap denote that the overlapping annotations called the same start and stop index for a given gene.

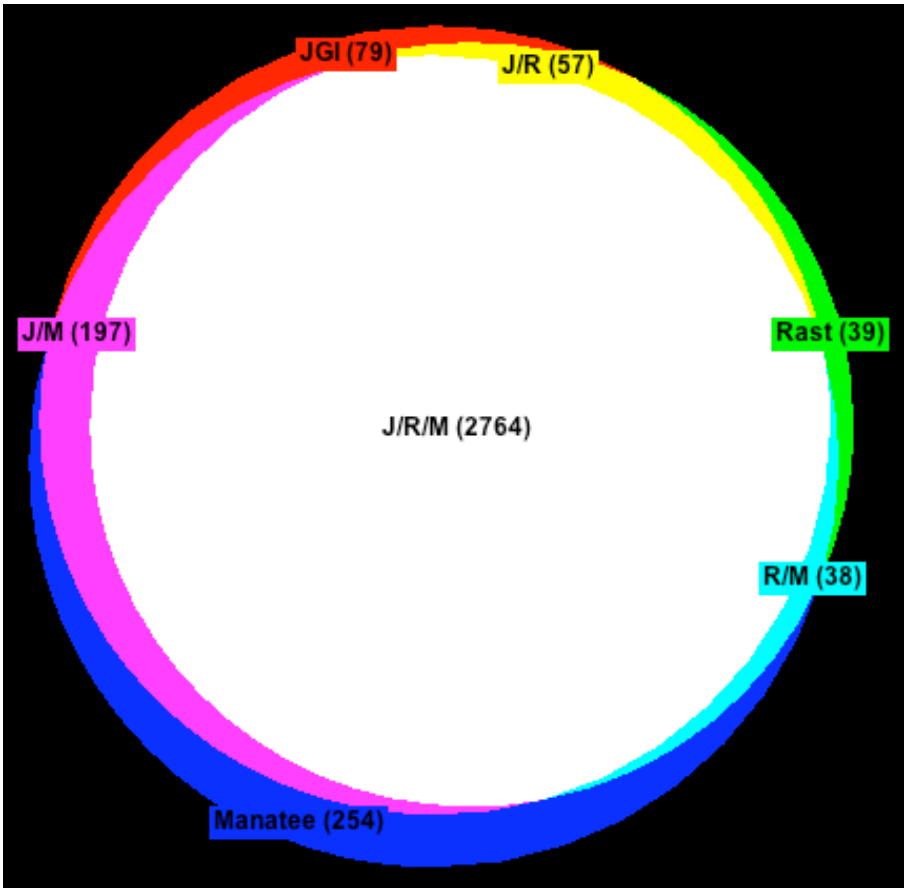

Figure 2 - Venn diagram showing the number of stop codon matches across the three annotations. Regions that overlap denote that the overlapping annotations called the same stop index and strand (+/-) for a given gene.

Looking closer at the gene calls made by each program, certain patterns emerged. We determined that Manatee called many more shorter genes than did either RAST or JGI (Figure 3). In an analysis of the predictions of about 3,000 genes, RAST more often called genes with an alternative start codon, identifying about 40% of start codons as something other than ATG, whereas JGI called alternative start codons about 15% of the time and Manatee 20% of the time (Figure 4). Interestingly, none of the databases ever called CTG as an alternative start codon.

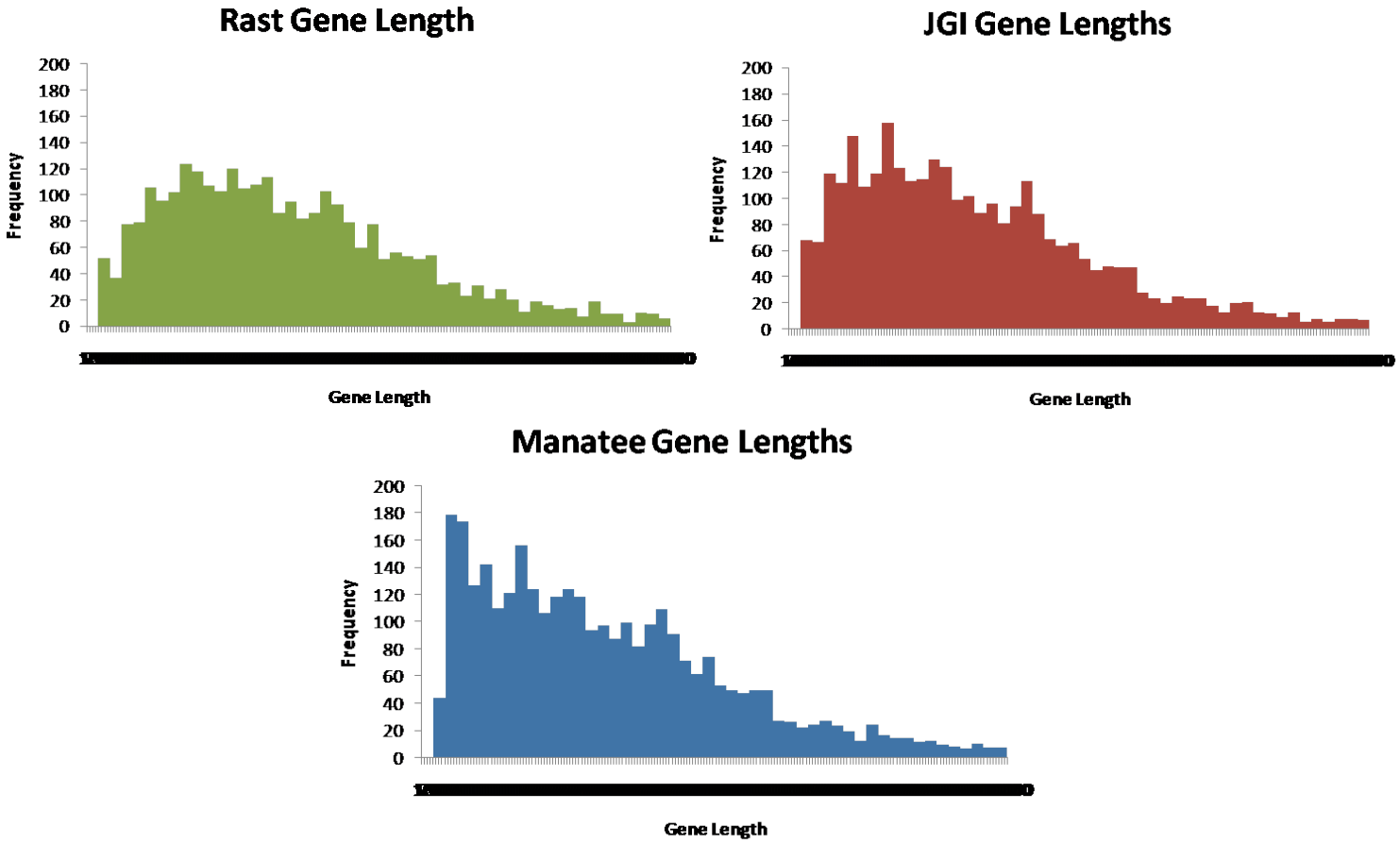

Figure 2 - Frequency of calls made as related to gene length in each of the three annotation engines.

Alternative Start Codons

Start Codon	JGI Predictions	RAST Predictions	Manatee Predictions
ATG	2604	1723	2562
Other	443	1128	646
Total	3047	2851	3208
Percentage Not ATG	14.3%	39.0%	19.9%

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

Figure 4 - Start codon predictions from each of the three annotations. The number given is the number of genes counted.

Figure 5 shows a comparison of average gene sizes under two conditions. We found that the average length of a gene when all three annotation engines called the same stop codon for that gene was very similar, just under 1,000 bases. This meant that the difference in

length was usually small if disagreement occurred between the annotations concerning the 5' end. However, when all three annotations called different stop codons, the length predictions of RAST were generally about twice as long as those of JGI and Manatee, a sizeable difference. So, if disagreement occurred between the annotations concerning the 3' end, the disagreements were usually significant between RAST and the other two databases.

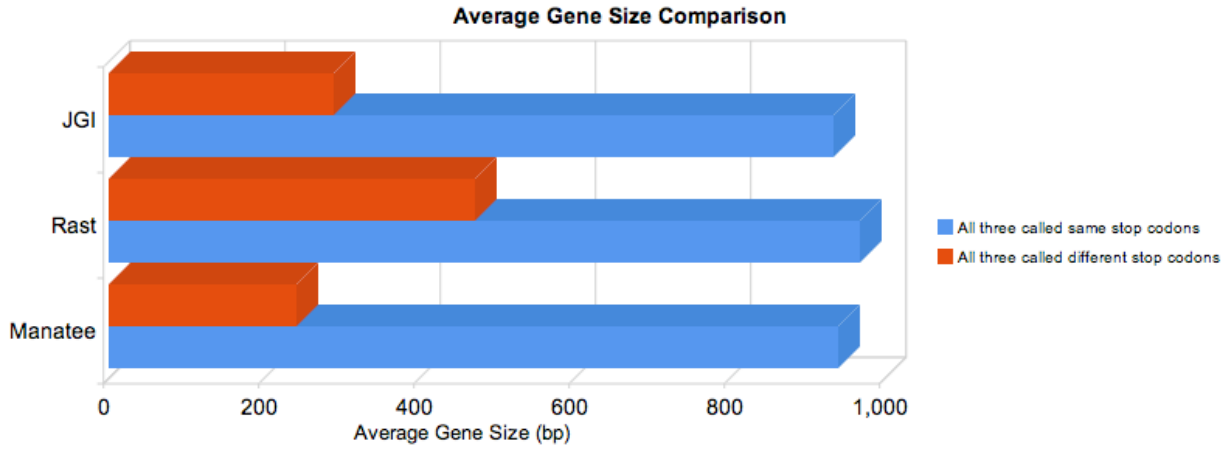

Figure 5 - Comparison of average gene sizes across the three annotations. The blue bars represent instances when all three engines called the same stop codons; the red ones represent instances where the they called different stop codons.

We analyzed the genome to try to find our genome-specific Shine-Dalgarno sequence and the complementary ribosomal binding sequence (RBS). We found a great diversity of RBSs, as shown in Figure 6; there did not seem to be a strong, consistent consensus sequence. However, the most common RBS was GGAGGTG, and the 16s rRNA subunit contains a perfect complement to that sequence (CCTCCAC, highlighted in Figure 7) near the 3' end, suggesting that CCTCCAC is the Shine-Dalgarno sequence and that GGAGGTG is the RBS. We also analyzed a number of gene-specific RBSs to find how strongly conserved each base in the RBS was. Figure 8 shows the results of that analysis; we only included RBSs that matched GGAGGTG or had only one base different, yielding about 250 genes in each annotation. The Gs were very strongly conserved, with the G in position 5 being most strongly conserved, while the A and T were less likely to be conserved. For those same genes in each database, we counted how far each RBS was from the start codon to see if they were the expected distance of 6-7 nucleotides upstream of the start codon, and we found that the majority of RBSs, about 65% in each annotation, were between 4 and 8 nucleotides upstream (Figure 9).

Sequence (7bp)	Frequency	ccggacc	41	ggaggga	33
ggaggtg	75	atcgaac	39	cgttttt	33
gatcgac	61	gggggtg	39	ggaacga	33
gaggtga	58	ctttttg	38	tcgaatc	33
cgatcga	53	gtccgga	38	acgtttt	32
cgaaacg	51	ccgaaac	38	gacgaaa	32
cggaggt	50	cggagga	37	gaaccga	32
cgacgga	49	cgacagt	37	tttatat	31
acggagg	47	gaccgaa	37	ttttgcc	31
gatcgaa	46	gacggag	37	accgatc	31
ccggagg	46	gaaacgc	37	tccgaac	31
cgaacga	46	cggaggg	36	tttatac	30
tcgatcg	45	cgacaga	35	gtttttg	30
ggatcga	45	aacgctt	35	ggggtga	30
ccgatcg	43	ggccgaa	35	cgaccca	30

Figure 6 – Frequency of occurrence of various RBSs, ranked in order from most frequent to least frequent.

>2500590728 HutaDRAFT_30940 16s rRNA 2397347..2398825(+) [Halorhabdus utahensis AX-2, DSM 12940]

TCCGGTTGATCCTGCCGGAGGCCATTGCTATCGGAGTCCGATTTAGCCAT
GCTAGTCGCACGGGTTTAGACCCGTGGCAAATAGCTCAGTAACACGTGGC
CAAACTACCCTGTGGACGGAAATAACCTCGGGAAACTGAGGCTAATGTCC
GATACGACTCGCCAGCTGGAGTGCGGCGAGTCGGAAACGTTGCGGCGCCA
CAGGATGTGGCTGCGGCCGATTAGGTAGACGGTGGGGTAACGGCCCACCG
TGCCCATAATCGGTACAGGTCATGAGAGTGAGAGCCTGGAGACGGTATCT
GAGACAAGATGCCGGGCCCTACGGGGCGCAGCAGGCGCGAAACCTTTACA
CTGCACGACAGTGCGATAGGGGGACTCCGAGTGCGAGGGCATATAGTCCT
CGCTTTTGTGTACCGTAAGGTGGTACAGGAATAAGGGCTGGGCAAGACCG
GTGCCAGCCGCCGCGGTAATACCGGCAGCCCGAGTGATGGCCGCTATTAT
TGGGCCTAAAGCGTCCGTAGCCGGCCAGACAAGTCTGTTGGGAAATCCAC
GCGCTCAACGCGTGGACGTCCGGCGGAAACTGTCTGGCTTGGGGCCGGAA
GATCTGAGGGGTACGTCCGGGGTAGGAGTGAAATCCCGTAATCCTGGACG
GACCGCCGGTGGCGAAAGCGCCTCAGAAAGACGGACCCGACGGTGAGGGA
CGAAAGCTAGGGTCTCGAACCGGATTAGATACCCGGGTAGTCCTAGCTGT
AAACGATGCTCGCTAGGTGTGCCGCAGGCTACGAGCCTGCGCTGTGCCGT
AGGGAAGCCGTGAAGCGAGCCGCCTGGGAAGTACGTCTGCAAGGATGAAA
CTTAAAGGAATTGGCGGGGGAGCACTACAACCGGAGGAGCCTGCGGTTTA
ATTGGACTCAACGCCGGACATCTCACCAGCACCGACAATGTGCAGTGAAG
GTCAGGTTGATGACCTTACTGGAGCCATTGAGAGGAGGTGCATGGCCGCC
GTCAGCTCGTACCGTGAGGCGTCCTGTTAAGTCAGGCAACGAGCGAGACC
CGCACTCTTAGTTGCCAGCAGCATCTTGCGATGGCTGGGTACACTAGGAG
GACTGCCGCTGCCAAAGCGGAGGAAGGAACGGGCAACGGTAGGTCAGTAT
GCCCCGAATGTGCTGGGCGACACGCGGGCTACAATGGCCGGGACAGTGGG
ACGCCAGTCCGAGAGGACGCGCTAATCCCCGAAACCCGGTCGTAGTTCGG
ATTGAGGGCTGAAACCCGCCCTCATGAAGCTGGATTCGGTAGTAATCGCG
TGTCAGAAGCGCGCGGTGAATCCGTCCCTGCTCCTTGCACACACCGCCCG
TCAAAGCACCCGAGTGGGGTCCGGATGAGGCCGTCATGCGACGGTCAAAT
CTGGGCTCCGCAAGGGGGCTTAAGTCGTAACAAGGTAGCCGTAGGGGAAT
CTGCGGCTGGAT**CACCTCC**TAACGATCGG

Figure 7 - The Shine Dalgarno sequence, highlighted in bold, in the sequence called as 16s rRNA.

Figure 8 - The ribosomal binding sequence (RBS) logo, generated by UC-Berkeley's WebLogo program. We used only those genes that contained the consensus sequence GGAGGTG or were only one base off from it, and included the 25 bp upstream of those genes; it does not illustrate the strength of consensus of the consensus sequence, only the relative strength of each letter in the consensus sequence.

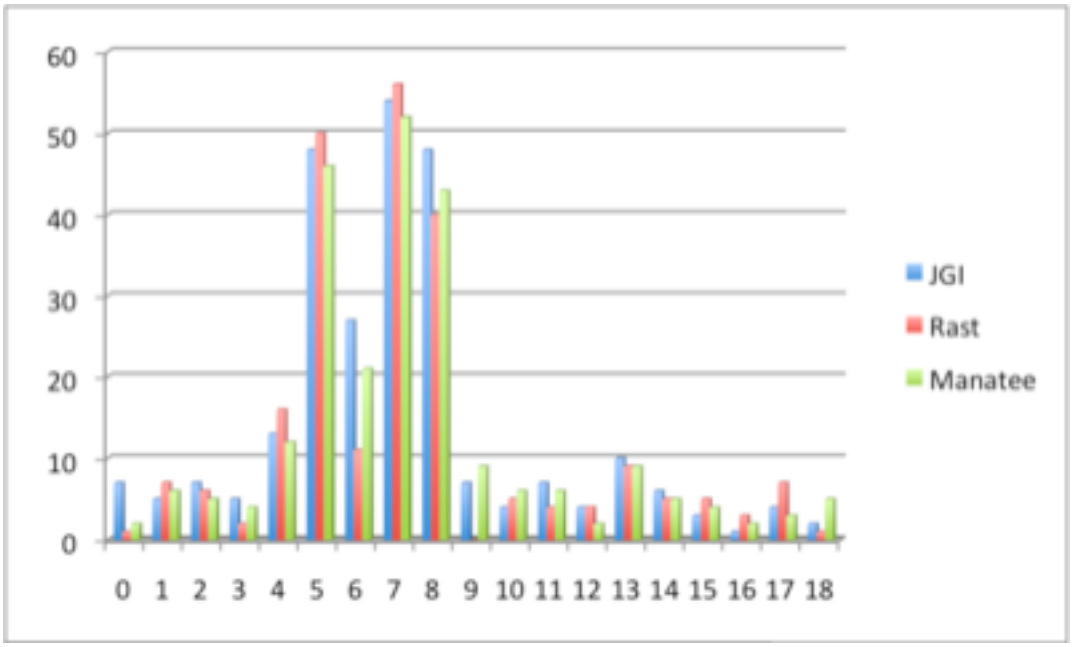

Figure 9 - The x-axis shows the position of the predicted RBS from the predicted start codon for genes in each annotation. The y-axis shows the frequency of occurrence. Most RBSs were between 5 and 8 base pairs from the start codon.

As a case study I investigated the mystery of the cation efflux system protein, providing a concrete example of some of the ways the three annotation engines differ. The same amino acid sequence was called in three different ways by the annotation engines: JGI called a “cation diffusion facilitatory family transporter” 303 amino acids long; RAST called a “cobalt/zinc/cadmium resistance protein“ 315 amino acids long; and Manatee called a

"cation efflux protein" 315 amino acids long. Upon investigation, I determined that the different names all describe the exact same kind of protein, and that RAST and Manatee seemed to have it right with the length, since they called the 5' end based on an alternative start codon, which matched orthologs in other species nearly perfectly. JGI called ATG as the start codon, making its 5' end truncated compared to the orthologs I looked at.

We also studied the phosphotransferase pathway to understand how helpful each annotation engine was for that analysis. No genes were predicted in SEED's KEGG pathway map prediction for phosphotransferases (Figure 9). After a text-based search of each database, I found one predicted phosphotransferase in SEED and JGI, and four in Manatee, but none of those matched any EC numbers from the phosphotransferase system described in the KEGG pathway map. There were many enzymes labeled as kinases, about 70 in each database, but I could identify none that fit in the pathway.

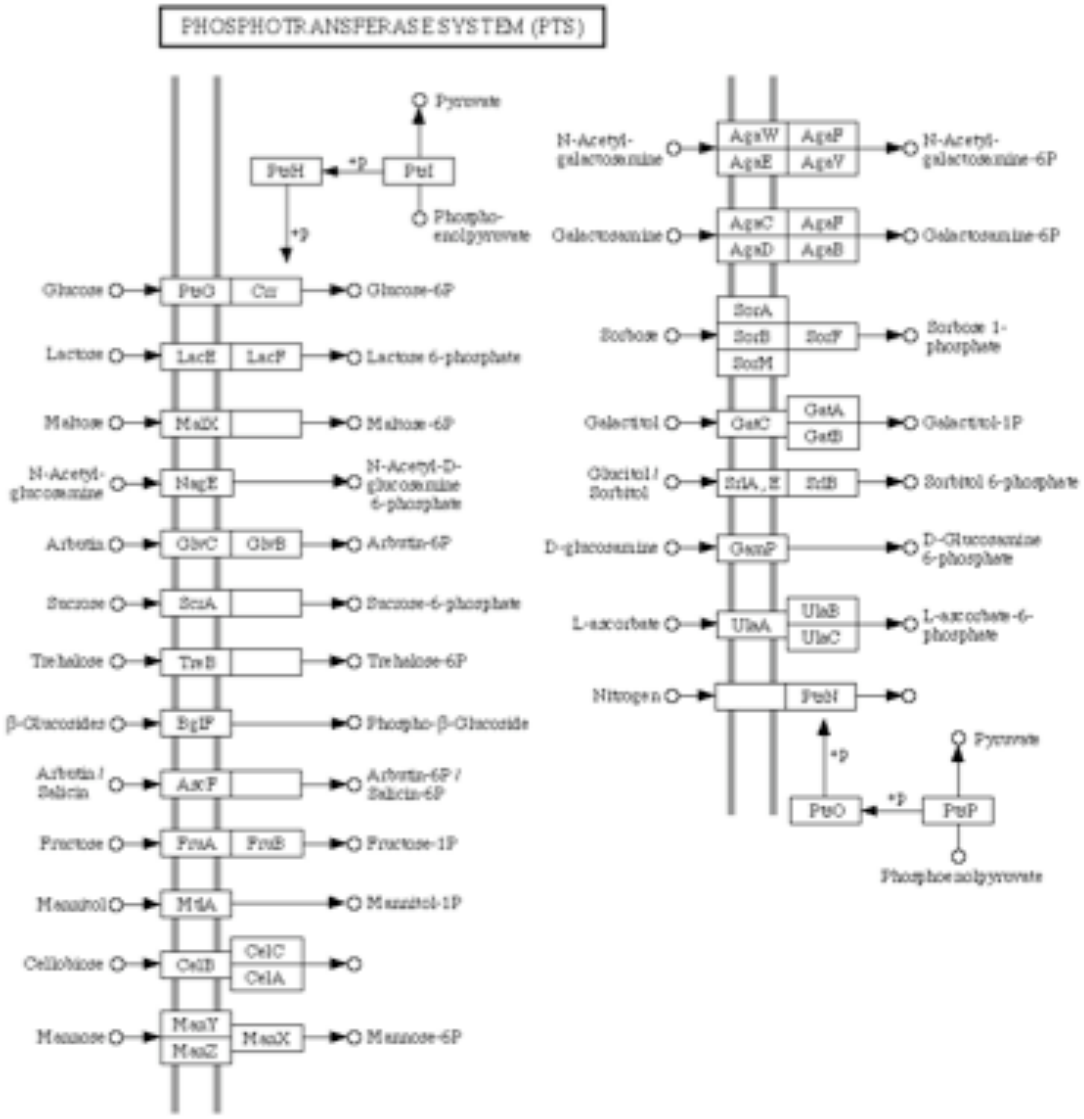

Figure 10 - The KEGG pathway map for *H. utahensis* as predicted by the SEED database; all of the boxes representing enzymes are white, which means that no matches are predicted.

Discussion

Only about half of the gene calls were exact matches in all 3 annotations; there is clearly significant difference in the ways that each annotation engine calls genes if they can only agree upon one half of the genes in an organism. One ought to expect this; if the annotations all agreed perfectly, there would be no need to have any more than one. None of them is yet perfect, and our analysis seeks to pinpoint the reasons behind the discrepancies and perhaps suggest which features seem to be less useful.

We saw from Figure 4 that RAST tended to call more alternative start codons. Manatee calls shorter ones, etc. This seemed to be the cause in Figure 1 of the lack of shared calls between RAST and either JGI or Manatee compared to the larger number shared between JGI and Manatee.

My case study on cation efflux system proteins showed the complications that arise when nomenclature is not standard in one database or between databases. A simplified and standard nomenclature for genes and proteins would make text-based searches more effective. The EC number system is very useful to that effect, as a standardized nomenclature, but even there the annotation engines have trouble calling all of the EC numbers actually present in an organism, and JGI, RAST, and Manatee are all limited by that.

In my analysis of the phosphotransferase, I can't claim to have performed a truly exhaustive investigation, yet I searched extensively, and still found nothing. This showcased for me the extent to which have to rely on the navigational features of the databases to learn about the genomes we annotate.

It would be very useful to conduct a further study comparing the fully curated, "finished" annotation of a genome to a first pass annotation through each of the automated annotation engines. This would provide further information on the relative efficacy of these three automated annotation engines in comparison to one another.

References

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. *The RAST Server: Rapid Annotations using Subsystems Technology.* BMC Genomics, 2008.

Hingamp P, Brochier C, Talla E, Gautheret D, Thieffry D, et al. (2008) Metagenome annotation using a distributed grid of undergraduate students. PLoS Biol 6(11): e296. doi:10.1371/journal.pbio.0060296

Wainø, M. & Ingvorsen, K. (2003). Production of ß-xylanase and ß-xylosidase by the extremely halophilic archaeon Halorhabdus utahensis. *International Journal of Systematic and Evolutionary Microbiology* 50, 183-190.

Acknowledgments

We thank Cheryl Kerfeld and Edwin Kim, JGI, for providing the sequences; Jonathan Eisen, UC Davis, for general advice; Gary Stormo, Washington University, for input on

Matt DeJongh, Hope College, for allowing us the use of the SEED annotation engine (supported in part by National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services (NIAD) under contract HHSN266200400042C) and for sharing several tools under development; Ramana Madupu, J. Craig Venter Institute, for allowing us the use of the JCVI annotation service and the automated Manatee annotation engine. We thank Kjeld Ingvorsen, Det Naturvidenskabelige Fakultet, Biologisk Institut, Aarhus Universitet, Denmark, for investigations *in vitro* to match ours *in silico*, and we also thank Chris Healey at Davidson College for ordering and growing *H. utahensis* locally.