Three-Way Comparison and Investigation of Annotated *Halorhabdus utahensis* Genome

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

We submitted the genomic sequence of halophilic archaeon *Halorhabdus utahensis* to be analyzed by three genome annotation services. We have examined the output from each service in a variety of ways in order to compare the methodology and effectiveness of the annotations, as well as to explore the genetics, pathways, and physiology of the previously un-annotated organism. We have found that the annotations services differ considerably in gene calls, features, and ease of use. We have also made important discoveries about *H. utahensis*, including the origin of replication, the consensus Shine-Dalgarno sequence, and the intron-containing tRNA-trp.

Introduction

The study of genomics has become increasingly important in today's world of science. More and more studies demand genomic data, often for multiple organisms living in one community or for organisms that cannot easily be grown in culture. Genomic methods and tools have surpassed manual human efforts as the amount of input data has increased by orders of magnitude. In order to benefit from the power of genomic sequencing, our tools must be streamlined, our databases must be consistent, and our knowledge of genomics must be strong. In the coming years, hundreds of genomes will be submitted to be sequenced and annotated. We must work to understand the necessary steps for improving the system of genomic annotation and the field of genomics as a whole. The goal of this study was to conduct a comparison of three annotation services: The Joint Genome Institute's Integrated Microbial Genome system, the National Microbial Pathogen Data Resource's Rapid Annotation using Subsystems Technology server, and the J. Craig Venter Institute Annotation Service. Additionally, our goal was to examine the *Halorhabdus utahensis* genome in order to understand the physiology and inner workings of a previously un-annotated organism. The halophilic archaeon *H. utahensis* offers a relatively simple genomic framework that encodes for a myriad of intriguing physiological traits. Isolated from the Great Salt Lake, Utah, *H. utahensis* grows optimally in 27% NaCl (Wainø, 2000). The genome is comprised of 3,129,561 base pairs that encode approximately 3,000 genes. Its small genomic size allows for manageable exploration of the genomic features as well as comparison between annotations. We have compared the annotations of three automated services as well as investigated the genetic workings of the organism. We have documented distinct differences in annotation output and review of the output. We have also discovered the origin of replication and the consensus Shine-Dalgarno sequence for this organism.

Materials and Methods

We received the *H. utahensis* strain AX-2 genome sequence in FASTA format from the Joint Genome Institute. The genome had been previously sequenced by JGI as part of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project. Whole-genome shotgun sequencing resulted in 5 contigs of varying sizes. The largest contig stretches 3,102,403 base pairs, representing over 99 percent of the genomic DNA. Four additional contigs measure 10,409, 9,346, 3,888, and 3,515 bases. JGI annotated the genome through their Integrated Microbial Genome Expert Review system (IMG/ER), and made the analysis publicly available on the IMG/EDU site (http://imgweb.jgi-psf.org/cgi-bin/img_edu_v260/main.cgi?section=TaxonDetail&page=taxonDetail &taxon_oid=2500575004).

We additionally submitted the *H. utahensis* genome sequence to two automated annotation services: The National Microbal Pathogen Data Resource's (NMPDR) Rapid Annotation using Subsystems Technology (RAST) server and the J. Craig Venter Institute (JCVI) Annotation Service. The RAST server provided a fully automated annotation of the genome, able to be browsed in a SEED-viewer environment (Aziz, 2008). The JCVI Annotation Service ran the genome through its Prokaryotic Annotation Pipeline and uploaded the output to Manatee, a web-based annotation tool and browser.

Web-based tools. We used numerous web-based tools in order to investigate the *H. utahensis* genome, as well as to compare the three annotation services. We utilized features included in the IMG, RAST, and Manatee browsers, including sequence exporters, open reading frame (ORF) visualizers, internal BLAST, and other search and comparison tools.

We also created software tools to facilitate our goals of exploring the genome and comparing the different annotations. We compiled these tools and other resources as part of the GCAT wiki page (http://gcat.davidson.edu/GcatWiki/index.php/Halorhabdus_utahensis_Genome). With one tool, we were able to search for an Enzyme Commision (EC) number in all three annotations with one keystroke. With another, we could BLAST multiple enzyme sequences against the genome by entering an EC number. Lastly, we enabled a text-based search of all three annotations' protein calls simultaneously.

Additionally, we used several outside web-based tools. We tinkered with tRNAscan-SE, a tool used by each of the three annotation services, in order to understand discrepancies in tRNA calls (http://lowelab.ucsc.edu/tRNAscan-SE/). We also utilized EMBOSS's "palindrome" tool to help locate the genome's origin of replication (http://emboss.bioinformatics.nl/cgi-bin/emboss/palindrome). Palindrome searches a DNA sequence to locate inverted repeats.

Results

Genome annotations provide a large quantity of data to analyze. In our analysis, we compared IMG, RAST, and JCVI annotations by examining gene calls, gene counts, start / stop sites, and EC numbers. We also examined evidence to determine the consensus ribosomal binding site and the DNA replication initiation site.

Comparison of IMG, RAST, and JCVI annotations. The first gene calls we examined were regions that code for ribosomal RNA (rRNA) and transfer RNA (tRNA). We found that IMG and RAST called three rRNA genes, whereas JCVI called only two.

IMG and RAST called identical 5s, 16s, and 23s rRNA strands. JCVI called the 5s and 16s rRNA, leaving the 23s rRNA missing. JCVI's 5s rRNA differed in start site from the other two annotations by one base. JCVI's 16s rRNA differed in start site by 18 bases and stop site by 986 bases (Figure 1).

IMG	DNA coordinates	Length
16s rRNA	2397347..2398825 (+)	1479 bp
23s rRNA	2399190..2402100 (+)	2911 bp
5s rRNA	2402216..2402338 (+)	123 bp
RAST		
16s rRNA	2397347.. 2398825 (+)	1479 bp
23s rRNA	2399190.. 2402100 (+)	2911 bp
5s rRNA	2402216.. 2402338 (+)	123 bp
JCVI		
16s rRNA	2397365.. 2397839 (+)	475 bp
5s rRNA	2402217.. 2402338 (+)	122 bp

Figure 1. Comparison of rRNA calls
Review of predicted coding regions for ribosomal RNA for each annotation service shows that IMG and RAST have identical calls, while JCVI fails to call 23s rRNA and predicts alternate start and stop sites.

After reviewing the tRNA gene calls made by each annotation service, we found that IMG and RAST had called 45 tRNAs, while JCVI had called 44. IMG and RAST made identical calls. We discovered that JCVI had failed to call an intron-containing tRNA-met, which IMG and RAST had called (tRNA-met, 1998587..1998721 (+), 135 bp). Further investigation revealed that none of the annotation services called a gene coding for tRNA-trp. Through additional searches, we found that the *H. utahensis* genome contains a gene coding for a tRNA intron endonuclease similar to that of *Halobacterium volcanii*, another halophilic archaeon (Thompson, 1988). We obtained the tRNA-trp sequence from *H. volcanii* from the Genomic tRNA Database, and BLASTed the sequence against the *H. utahensis* genome (http://lowelab.ucsc.edu/GtRNAdb/). The search revealed a tRNA-trp in the genome with 90 percent identity, containing a 103-base intron (tRNA-trp, 465601..465777 (-), 177 bp).

For less highly conserved genes, the annotation systems differed more often in their gene calls. Number of predicted genes ranged from 2,898 to 3,254, and the average gene length was between 845 and 942 base pairs. JCVI predicted the highest number of genes, followed by IMG, then RAST. However, on average RAST called considerably longer genes than IMG or JCVI (Figure 2). JCVI had the shortest average gene length, and called considerably more short genes than IMG or RAST (Figure 3).

Annotation	Genes	Mean	Median	Minimum	Maximum
IMG	3097	869.9 bp	728 bp	70 bp	7130 bp
RAST	2898	941.8	801.5	70	100001
JCVI	3254	844.9	692	73	100001

Figure 2. Comparison of descriptive statistics
Mean, median, minimum, and maximum gene lengths of the total predicted coding regions illustrate differences in gene calls between the annotations.

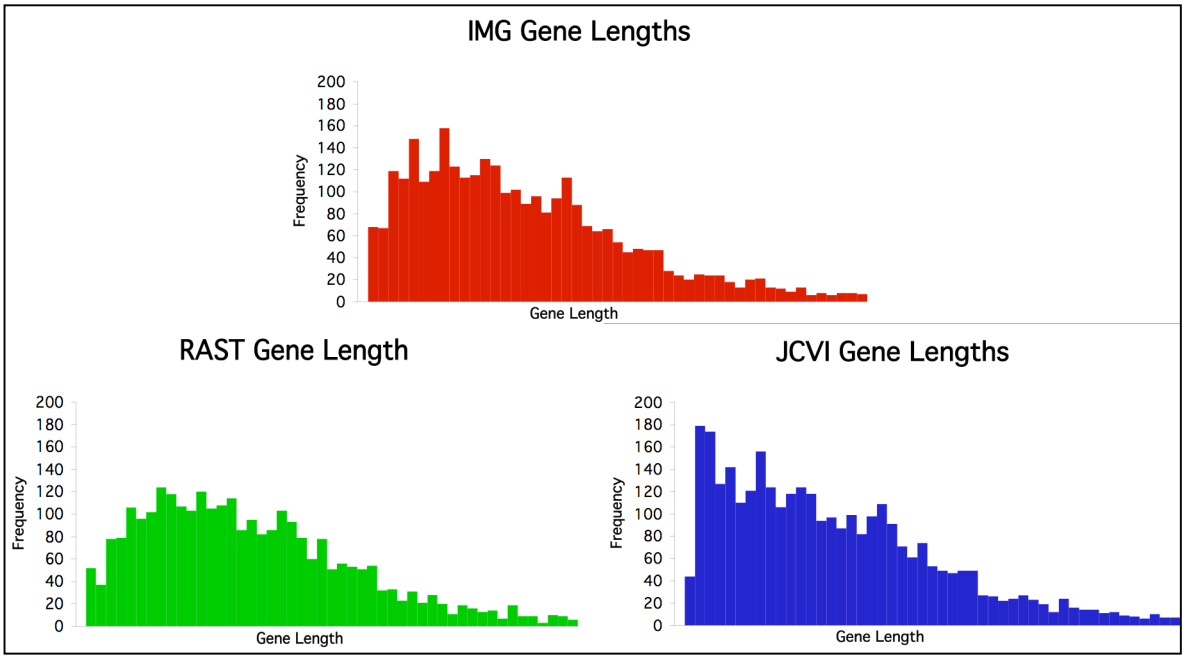

Figure 3. Comparison of gene length frequencies
Histograms displaying gene length illustrate similarities and differences between annotations. It is apparent that JCVI has a higher frequency of short genes called than IMG or RAST.

Discrepancies in gene calling can be categorized as differences in start site or differences in reading frame. Figure 4 shows that the annotations agreed on reading frames much more often than they agreed on start site. JCVI had the largest number of genes with distinct stop sites. However, 80 percent of all predicted protein coding genes shared stop sites with both other annotations. When comparing exact matches, it is apparent that IMG and JCVI shared a great deal more exact gene calls with one another than with RAST. RAST had the largest number of genes that differed from the other two annotations. For predicted protein coding genes from all annotations, only 29.6 percent were identical across the three annotations.

To further analyze differences in start site, we tabulated the start codon for each predicted gene. ATG was the most common start codon across all annotations, accounting for 75.7 percent of the starts. RAST contained the largest proportion of alternative start codons, with 39.0 percent of the genes predicted to begin with a codon other than ATG. JCVI and IMG had considerably lower alternative start usage, with 14.3 and 19.9 percent, respectively (Figure 5).

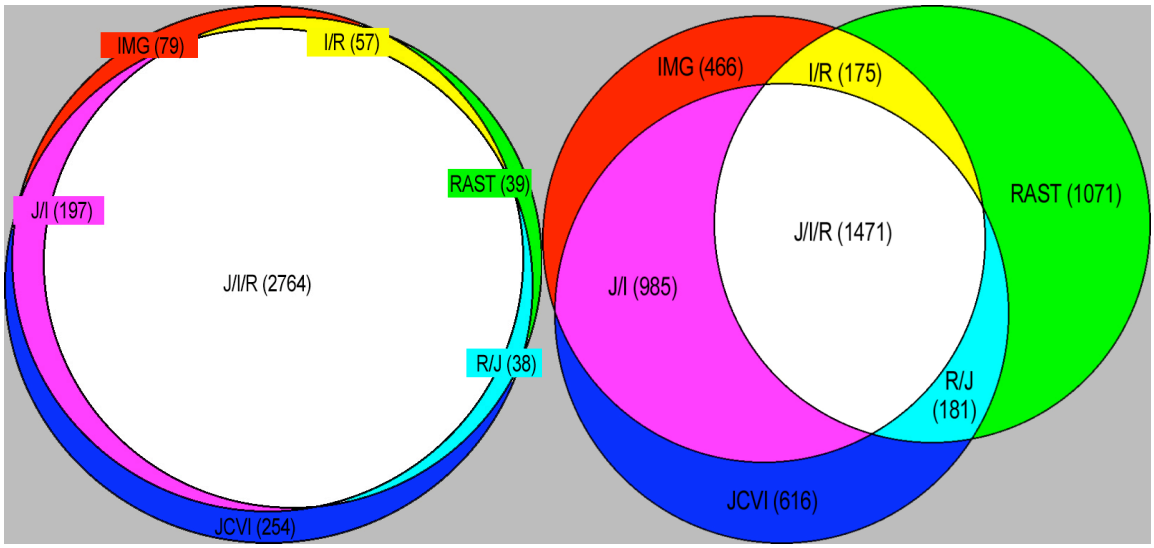

Figure 4. Venn diagrams of gene predictions

(A) The diagram to the left shows the number of predicted protein coding genes that share stop sites with the other annotations. Overlapping regions indicate genes having same stop site between annotations. (B) The diagram to the right shows the number of predicted protein coding genes that share start site and stop site with the other annotations. Overlapping regions indicate genes having exact matches between annotations.

Annotation	Genes	ATG Start	Other Start	% Not ATG
IMG	3047	2604	443	14.3%
RAST	2851	1723	1128	39.0%
JCVI	3208	2562	646	19.9%

Figure 5. Comparison of start codons

Analysis of predicted protein coding genes displays incidence of ATG and alternative start codons for each annotation.

Figure 6 shows that for genes that shared stop codons with both other annotations, genes called by RAST had the longest average length. The average length for RAST was 967 base pairs, in comparison to 940 for JCVI and 934 for IMG. Also, for genes unique to one annotation, RAST calls had the longest average gene length by a wide margin, followed by IMG, then JCVI (Figure 6).

Gene lengths of hypothetical proteins follow the same pattern as genes with unique stop codons. Each annotation service labeled slightly over 1,000 genes as hypothetical or unknown. For these genes, RAST's hypothetical protein genes were the longest, followed by IMG, then JCVI. In all annotations, the average hypothetical protein was considerably shorter in length than the average length of all predicted proteins. RAST's average hypothetical protein was 23 percent shorter than RAST's total average gene length. IMG's average was 28 percent shorter than its own total average gene length. Finally, JCVI's average was 39 percent shorter than its own total average gene length. JCVI's hypothetical proteins are shortest both in raw length and as a percentage of the total mean (Figure 7).

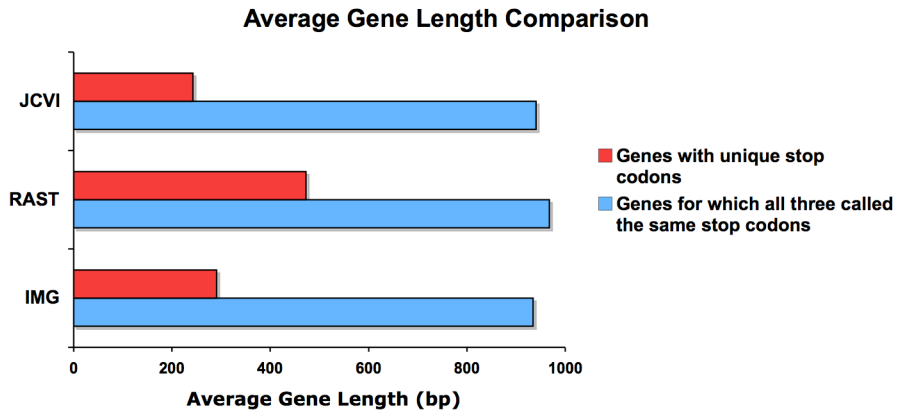

Figure 6. Comparison of average gene length
Illustrates average gene length for two categories. Red bars represent the average length of genes from each annotation that have distinct stop codons. Blue bars represent the average length of genes that have a common stop codon across the three annotations.

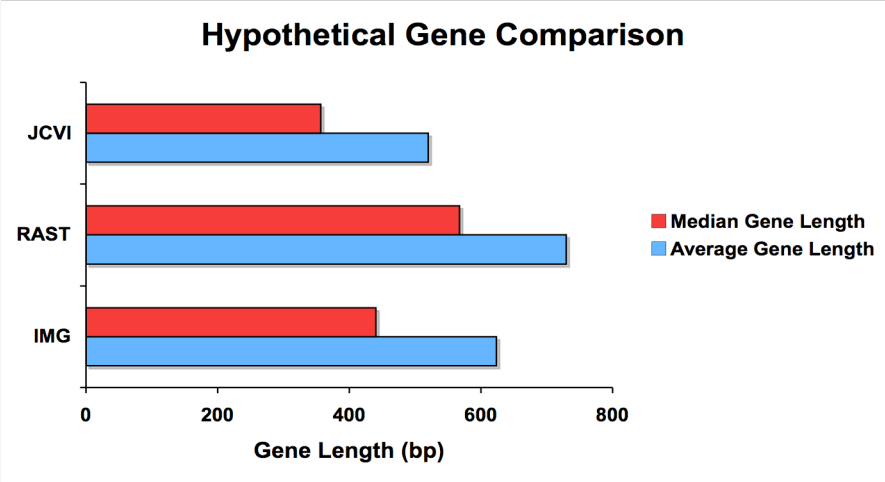

Figure 7. Comparison of hypothetical protein length
Analysis of genes called as "hypothetical protein," "conserved hypothetical protein," or "unknown" shows that these calls differ greatly between annotations. Red bars represent the median length of hypothetical genes from each annotation. Blue bars represent the average length of hypothetical genes.

The addition of EC numbers to predicted genes provides a specific, universal classification for enzymes. EC numbers aid the organization of genes into pathways and subsystems. We tallied the predicted genes in each annotation that had been labeled with either full or partial EC numbers. We found that RAST assigned 498 (17.5 percent) of its genes with an EC number. JCVI assigned EC numbers to 485 (15.1 percent) of its genes. Finally, IMG assigned only 196, or 6.4 percent of its genes with an EC number.

Generation of a consensus RBS. Studies indicate that ribosomes are often recruited for translation by a sequence closely upstream from the coding region, known as the Shine-Dalgarno (SD) sequence. Because SD sequences often reside several bases upstream of the start codon, finding a species-specific SD sequence could be useful in determining or altering start sites.

In order to generate a consensus ribosomal binding site (RBS), we analyzed the 50 bases upstream of 19 polymerases and 32 large subunit ribosomal proteins. It was

beneficial to use these genes because they are highly conserved and highly expressed, which aids in the conservation of the upstream regions. We recorded and aligned recurring motifs through manual curation. We found 32 possible RBS sequences out of the 51 genes. Then we created a position-base frequency plot in order to condense the data into a consensus sequence (Figure 8). We found that the consensus RBS was a SD sequence, GGAGGT, found 7 to 11 bases upstream of the start codon. We later used bioinformatics tools to generate a more statistically sound consensus Shine-Dalgarno sequence. This will be discussed at greater length in a colleague's paper.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A	10	8	4	0	20	0	0	6
C	11	13	11	0	5	0	0	5
G	11	3	15	32	3	32	32	5
T	0	8	2	0	4	0	0	16
	not T	c	g	G	A	G	G	T

Figure 8. Shine-Dalgarno position-base frequency plot
The position-base frequency plot facilitates the generation of a consensus SD sequence. The position (0-7) of each base (A/C/G/T) in each preliminary SD sequence is recorded. Bases with high frequency at a certain position are given consensus base status. Approximate consensus sequence is GGAGGT.

Location of replication origin. We located the DNA replication initiation site by searching for evidence outlined in several papers concerning archaeal origins of replication (Burquist, 2003) (Duggin, 2006). First, we located genes in the *H. utahensis* genome that code for the archaeal equivalent of an Origin Recognition Complex subunit (ORC) and a cell division control protein (Cdc6). These ORC/Cdc6 homologues are good indicators because they often lie in close proximity to the replication origin. We discovered five ORC/Cdc6 homologues in the genome. Due to the proximity of DNA polymerase, helicase, and other replication factor genes, we examined the area surrounding ORC/Cdc6 3 (2324949..2326724 (-)) (Figure 9).

Upon closer investigation, we found supporting evidence that this region contained the origin of replication. We discovered a non-coding, AT-rich, 1,000 base pair region upstream of the ORC/Cdc6 3 gene (2326724..2327725). The region possesses a 49 percent GC content in comparison to a 63 percent genome average. It also contains a pair of 28-base inverted repeats, which form a transcription factor binding site when coiled (2327117..2327142 (+), 2327719..2327745 (-)) (Grabowski, 2003). Other supporting evidence includes opposite-facing genes and a local minimum in cumulative GC skew. This brought us to conclude that the origin of replication for *H. utahensis* lies approximately at base 2,327,225 of the primary contig.

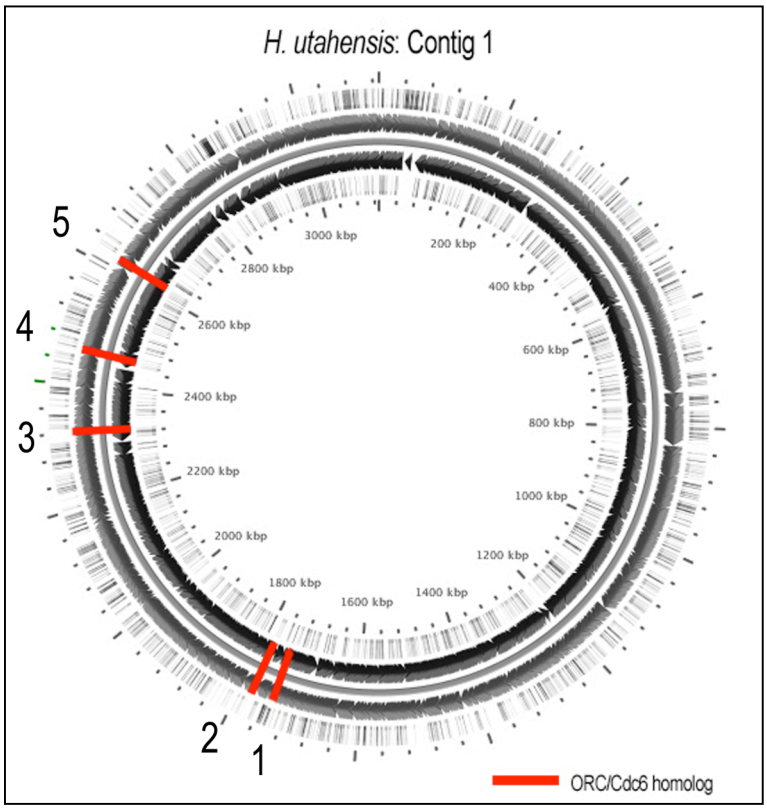

Figure 9. *H. utahensis* primary contig and ORC/Cdc6 genes
Circular display of the largest contig of the *H. utahensis* genomic sequence. The contig begins at the top and wraps clockwise. The red bars illustrate the location of ORC/Cdc6 homologs. The ORC/Cdc6 gene numbered 3 lies near the origin of replication, approximately at 2,327,225 bases.

Discussion

rRNA and tRNA. It was surprising to find that JCVI had failed to locate one rRNA and had badly truncated another. rRNA genes are among the most conserved regions of DNA in archaea and bacteria. The reason for JCVI's annotation errors may have been a difference in tools used to find the rRNAs. JCVI used BLAST and Rfam to locate rRNAs, whereas IMG used an IMG RNA database and RAST used a script by Niels Larsen (Aziz, 2003).

JCVI also missed a tRNA-met where the other two annotations found it. This discovery is interesting because all three annotation services use a program called tRNAscan-SE to locate tRNAs. For this omission to occur, JCVI may have lowered the default cutoff for tRNA length in tRNAscan-SE, which may have passed over the 135 base pair tRNA-met (Lowe, 1997). In the case of the missing tRNA-trp, none of the annotations called the gene. tRNAscan-SE most likely missed the tRNA-trp because the program overlooks potential tRNAs that contain an intron greater than 80 bases.

Gene lengths, starts, and stops. The patterns that emerge from average length, start site and stop site agreement, and start codon data concern RAST and JCVI. RAST genes were longer than the others, on average. This was, in part, due to the increased calling of alternative start sites by RAST. In fact, there was a direct correlation between

alternate start codon use and average gene lengths between the annotations. Additionally, JCVI had more short genes than RAST or IMG. This may have been a result of the service calling many short, hypothetical protein genes. The difference between gene calls of JCVI and RAST was intriguing because of their use of similar annotation tools. Both used the Gene Locator and Interpolated Markov Modeler (GLIMMER) tool for the first pass at genes (Aziz, 2003). The differences may come about in the training set given to GLIMMER before the genome runthrough, which is not consistent between annotation services. Significant changes also may come about during additional analysis through different tools and searches of databases.

Ease and functionality of browsers. Each web-based viewer offered helpful tools and features for research and analysis. The ease of exporting DNA or amino acid sequences for genes made IMG/EDU a valuable resource. Also, the ability to perform a text search of 57 annotated archaeal genomes and then BLAST a selected gene against *H. utahensis* was a powerful tool. However, the inability to BLAST any DNA or amino acid sequence hindered us from finding the tRNA-trp in the genome. For that, we were forced to use the RAST SEED-viewer browser. The quick and easy BLAST against *H. utahensis* genome was the most-used feature. Also, the ease of sorting and searching the entire predicted gene list was helpful. JCVI's Manatee browser had a feature that grouped certain genes together based on function. This greatly aided our search for the origin of replication by compiling the many genes involved in the process into one page. However, Manatee was sluggish, error-ridden, and not intuitively designed. The site was more of a tool for manual annotation rather than a streamlined browser for data and analysis.

Future implications. Our exploration of the *Halorhabdus utahensis* genome as well as side-by-side comparison of three annotation services and browsers provide numerous possibilities for future research. Laboratory research could be carried out using live cultures of *H. utahensis*. Such experiments would provide additional evidence to support hypothesis about pathways and genes in the organism.

Also, work could be done to integrate our ideas and tools into existing or future annotation systems. The scientific community would benefit from increased consistency, interconnectedness, and ease of use in genomic annotation.

Acknowledgements

We thank Cheryl Kerfeld and Edwin Kim, JGI
Jonathan Eisen, UC Davis
Gary Stormo, Washington University
Matt DeJongh, Hope College for SEED/RAST
Ramana Madupu, J. Craig Venter Institute for Manatee
Kjeld Ingvorsen, Det Naturvidenskabelige Fakultet Biologisk Institut,
Aarhus Universitet, Denmark
Chris Healey at Davidson College for ordering and growing *H. utahensis* locally.

References

Aziz RK, Bartels D, Best AA, Dejongh M, Disz T, et al (2008) The RAST Server: Rapid Annotation using Subsystems Technology. BMC Genomics 9: 75.

Barry ER, Bell SD (2006) DNA Replication in the Archaea. Microbiol Mol Biol Rev 70: 876-887.

Berquist BR, DasSarma S (2003) An Archaeal Chromosomal Autonomous Replicating Sequence Element from an Extreme Halophile, *Halobacterium* sp. Strain NRC-1. Joural of Bacteriology 185: 5959-5966.

Duggin IG, Bell SD (2006) The Chromosome Replication Machinery of the Archaeon *Sulfolobus solfataricus*. J. Biol. Chem. 281: 15029-15302.

Grabowski B, Kelman Z (2003) Archaeal DNA Replication: Eukaryal Proteins in a Bacterial Context. Annual Review of Microbiology 57: 487-516.

Ivanova NN, Mavromatis K, Chen IA, Markowitz VM, Kyrpides NC. Standard Operating Procedure for the Annotations of Genomes and Metagenomes submitted to the Integral Microbial Genomes Expert Review (IMG-ER) System.

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25: 955-964.

Markowitz VM, Korzeniewski K, Palaniappan K, Szeto E, Werner G, et al (2006) The Integrated Microbial Genomes (IMG) System. Nucleic Acids Res. 34: D344-D348.

Thompson LD, Daniels CJ (1988) A tRNA-trp Intron Endonuclease from *Halobacterium Volcanii*. The Journal of Biological Chemistry 263: 17951-17959.

Wainø M, Tindall BJ, 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 Systematic and Evolutionary Microbiology 50: 183-190.