An Examination of the Discrepancies Between Three Genome Annotations of *Halorhabdus utahensis*

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

[1]Davidson College Biology Department
[2]Davidson College Math Department

Abstract

The field of genomics holds incredible potential, and it is imperative that students are exposed to the opportunities that the field offers to prepare the future generations of scientists and leaders. However, as a discipline that continues to develop, genomics still provides many issues that must be addressed, particularly with regards to the computerized and automated tools used to analyze entire genomes. In this study, we address both of these issues in as a class of undergraduate students analyzes the accuracy and reliability of three automatic annotation engines via the annotation of *Halorhabdus utahensis*, a halophilic archaeon. We conclude that the many discrepancies between the databases must be remedied by universal standardization of annotation software, in addition to continuing to improve the automated tools, and that multiple annotations of a genome should be compared to provide the most accurate analysis.

Introduction

The field of genomics continues to grow in importance in the scientific world. The examination of entire genomes through the use of mathematical tools and computer software unlocks a world of powerful possibilities. The applicability of the field to pressing health, environmental, and ecological issues means that its applications will transform the manner in which we interact with the biological world. Genomics has begun to play a key role in the field of nutrition research [1-2]; has applications in conjunction with biotechnology to improve the yield and quality of the food supply to match population growth [3]; holds great promise in the field of Phylogenomics to further the study of evolution [4]; and has the potential to be used to help equalize health between the developing and the developed worlds [5]. With so many applications and possibilities, genomics resembles a proverbial elephant in the room that is impossible to ignore.

The current crop of undergraduate students will need to be well-versed in the tools and methodology of genomics. Fortunately, genomics is not only an extremely relevant field, but it remains inexpensive and easy to tailor a variety of experience levels. The most effective manner in which to delve into this practical and exciting field is to have students apply their knowledge through first-hand laboratory experience [6]. As a result of the application of genomics in a classroom setting, students have not only benefited from a greater understanding of genomics but have also grown to appreciate the research process itself [7].

With this in mind, the Davidson College Laboratory Methods in Genomics class undertook an analysis of a member of the Archaea, *Halorhabdus utahensis.* This archaeon is an extremely halophilic microbe that found in the Great Salt Lake in Utah, and was first described

and classified in 2000 as a new species of a new genus [8]. The species exhibits a number of surprising features. The cells are highly pleomorphic and have a variety of different forms, although a rod-shaped form was found most commonly in younger cultures. They were capable of growing in a range from 9% NaCl concentration to 30% concentration, the point of saturation of the medium at 30°C. Likewise, the cells could also withstand a wide range of temperature and pH levels (17-55° C at 27% NaCl concentration and a pH from 5.5-8.5 at 30°C at 27% NaCl concentration). Surprisingly, the species grew only a few sugars, including glucose, xylose, and fructose, and amino acids, alcohols, and other carbon sources that were tested did not result in cell growth.

We began with a publically available annotation of the *H. utahensis* genome provided by the Department of Energy's Joint Genome Institute (JGI). Our original goal was to examine the accuracy of annotation of the specific organism, primarily as a learning exercise. We soon discovered, however, that the massive volume of genomics data provides a challenge to any attempt to validate an entire genome by hand; the total information contained in the databases contains far more information than is humanly possible to examine without effective tools [9], and the quality of those tools determines the worth of the data analysis. Therefore, we soon altered our purpose and began to compare three different annotations of the organism's genome and to verify the validity of the computer-automated annotation of the genome.

To our knowledge, there has never been a study that compared the different annotations of a particular genome provided by different databases. We intend to provide an answer to the question of whether these databases each have their own particular tendencies that tend to skew their protein calls and would affect the overall quality of the annotation. Our study raises the question of what determines the "correct" annotation of the genome; if there are discrepancies between the annotations, we intend to address how the scientific community could deal with the various interpretations of the genome.

Methods and Materials

We examined the *H. utahensis* genome using three distinct annotation engines, all of which approach the annotation with different methods and guiding principles. In JGI's Integrated Microbial Genomes (IMG) database, all new genomes are subjected to a validation process that corrects start codons and protein coding sequences that overlap in addition to checking for any genes or pseudogenes that have not been called [10]. Furthermore, to prevent discrepancies in protein calls, the IMG database also assigns "IMG terms" that designate general function to organize various genes by function into larger categories of "IMG pathways." JGI experts define these terms and pathways from specific genomes, and the terms are subsequently applied to other organisms to help ensure accuracy throughout the database [11].

The genome was also annotated by Rapid Annotation using Subsystem Technology (RAST), accessed via the SEED database. RAST approaches the annotation of a genome from a more holistic approach, based on the premise that genome analysis and modeling is both more accurate and complete when individual "subsystems" of an organism form the basis of annotation rather than attempts to apply functions to individual genes [12]. The annotation engine has been designed to allow more informed consideration of the functions of various genes within the context of its role in a subsystem [13].

Finally, the genome was also annotated by the J. Craig Venter Institute using its Manatee database. Manatee itself represents a compilation of data from various annotation tools that utilize the same annotation techniques to create a complete analysis of the entire genome [14]. The database is designed, however, to allow the user to edit the annotation and to store evidence for annotation specifics to allow improvement of the automatic curation [15].

We began by examining the JGI annotation of the genome. Through the use of several other websites, we verified the accuracy of various gene calls made by the IMG database. A primary instrument in this investigation was the National Center for Biotechnology Information's (NCBI) BLAST tool, which proved extremely valuable in the comparison of different nucleotide and amino acid sequences and also identification and verification of the identity of various protein calls from the annotations. Likewise, ExPASY's Enzyme site, a database of information on various enzymes, proved a valuable source of nucleotide and amino acid sequences and function of various enzymes. NCBI's Conserved Domains Database (CDD) also facilitated comparison of a sequence's predicted protein function to the COG (Clusters of Orthologous Groups) that the sequence likely belonged too; along with the Protein Data Bank (PDB) and the Sanger Institute's PFAM, this allowed verification of protein function. Finally, the Center for Biological Sequence Analysis' SignalP tool was used in conjunction with PSORT to analyze the likely location of protein within the cell to verify the accuracy of protein calls.

We also inspected the enzymatic and biochemical pathways of the organism to facilitate greater comparison of the annotations and verify their accuracy. The Kyoto Encyclopedia of Genes and Genome's (KEGG) Pathway tool, which did not contain data *H. utahensis* but which did include *Halobacterium salinarium*, the closest relative of *H. utahensis* [8], was used in conjunction with the SEED's colorized KEGG pathway diagrams. These tools allowed us to examine the predictions of different biochemical and enzymatic pathways contained within the *Halorhabdus utahensis* genome. The SEED's set of species-specific pathways was particularly useful in determining whether the organism could theoretically accomplish certain enzymatic tasks. We verified various pathways by searching the other annotation engine databases for the enzymes that the SEED marked as present or absent.

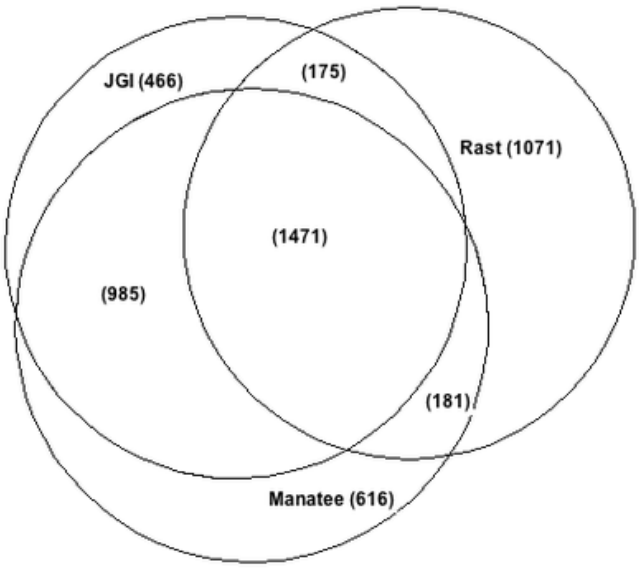

Figure 1. Exact gene matches across the 3 annotations. Regions that overlap denote that the overlapping annotations called the same start and stop index for a given gene.

Furthermore, we utilized a number of tools created by students to verify pathway predictions. Among this software was a program that searches all three genome annotations to display whether a particular E.C. number is called in any of the databases [16]; a program that obtains known amino acid sequences for a requested EC number and will BLAST those sequences against *H. utahensis* protein calls to determine whether the organism produces the particular enzyme [17]; and a tool that allows the user to perform an exact-hit search of the JGI, RAST, and Manatee annotations of the genome [18]. These programs allowed us to accurately pinpoint specific enzymes to determine whether the annotations agreed in their protein calls.

We compiled the data and tools that students created, including tutorials explaining the functioning of various databases utilized throughout the project, into a public wiki webpage [19].

Results

General Annotation Differences

The JGI database, IMG, called 3126 total genes from the *Halorhabdus utahensis* genome. Manatee and RAST predicted 3253 genes and 2915 genes, respectively. IMG determined that of the total predicted genes, 3076 had a likely function, while Manatee assigned a function to 1717 of the called genes. The SEED did not list the number of predicted genes with function. Of the total number of called genes in each annotation, only a certain percentage were exact matches with both the same start and stop index of another gene in the other two annotations (Figure 1). Of the total called genes, 1471 genes were predicted with exactly the same start and stop codon. A number of other genes matched a gene in one of the other database annotations but not the third; this overlap of databases was greatest between IMG and Manatee, which share 2456 exact gene matches, 985 of which were not predicted by RAST. The fewest matches were between RAST and IMG, which shared only 1646 identical protein calls, 175 of which were not called by Manatee as well.

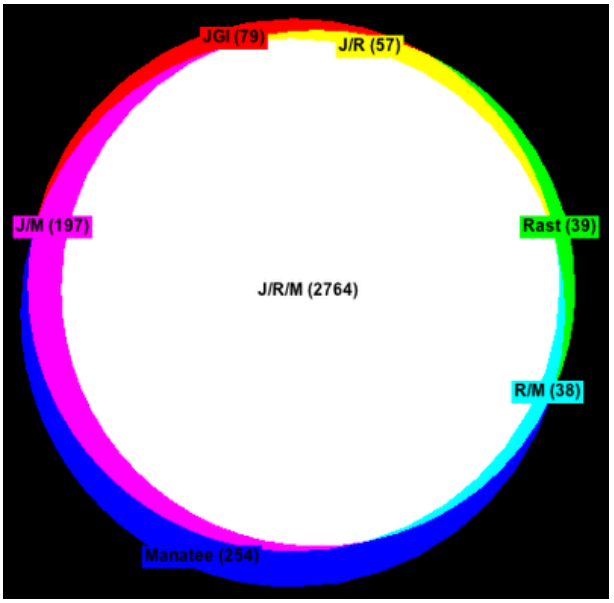

Figure 2. Stop codon matches across the 3 annotations. Regions that overlap denote that the overlapping annotations called the same stop index and strand (+/-) for a given gene.

The variation in gene calls can be further scrutinized through an examination of particular stop codons that the annotation engines called. There is greater consistency in the predicted stop codons than genes that are exact matches (Figure 2). While only 1471 gene calls matched exactly in all three databases, there were 2764 instances in which the databases predicted the same stop codon for a gene on the same nucleotide strand. Therefore, the difference in predicted genes was more clearly reflected in the variance of the start codons called by the software; the data suggests that the start codons must have been different in 1293 gene calls, even though the databases predicted the same stop codon. Thus, the greatest variance in the gene calls between the annotation engines was in their disagreement over start codon predictions.

This general trend was illustrated by the specific example of the gene that the JGI database labeled as 2300587691, a predicted glycoside hydrolase, with the coordinates (69942...72866). All three databases predicted the same stop codon (TAA at index 72866), but different start codons and start index coordinates (Table 1). While the JGI and Manatee annotations both determined that the gene began with an ATG start codon, the SEED predicted that it instead began with a GTG codon. Furthermore, although the IMG and Manatee databases both predicted an ATG

Table 1. Various annotations of gene labeled 2300587691 by IMG.

Annotation	Predicted Coordinate	Predicted Start Codon
IMG	69942...72866	ATG
Manatee	69882...72866	ATG
SEED	69912...72866	GTG

Table 2. Statistics for the comparison of gene length predicted by the three annotations.

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

codon, the annotation engines chose different ATG codons at different locations to mark the beginning of the gene.

The difference in the start and stop indexes of the gene calls was reflected also in the length of the genes predicted by each database. The average length of predicted genes in the RAST annotation was longer than that of both Manatee and JGI by an average of 96.9 and 71.9 bp, respectively, while Manatee gene predictions had a shorter average length then the gene calls of the other annotation engines (Table 2). While all three annotations do predict a varied range of gene lengths, they each call a greater percentage of genes of a certain length (Figure 3). Thus, the variation in start codon calls the SEED displayed a greater tendency than the other two databases to call alternative start codons that resulted in longer gene sequences with different indices.

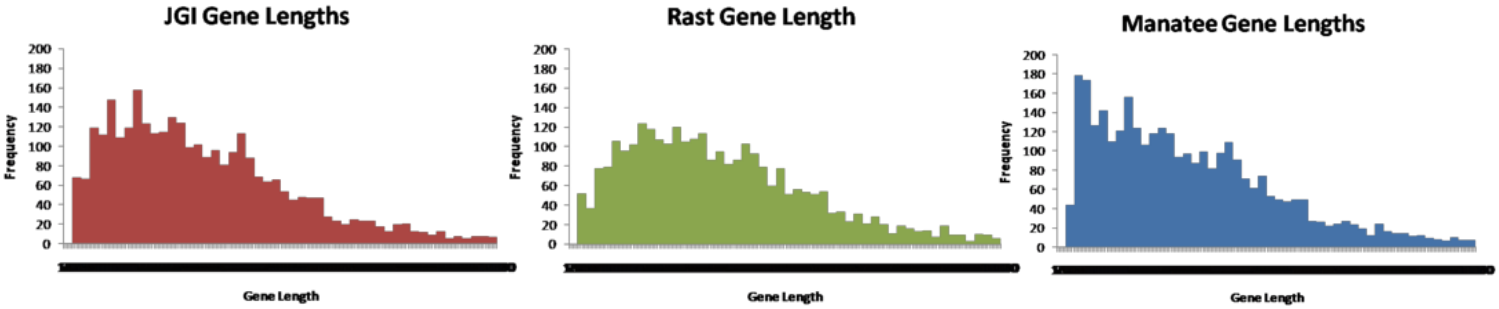

Figure 3. The ranges of gene lengths called by each annotation engine. Manatee predicts the greatest number of genes with a shorter length, while RAST predicts a greater number of longer genes.

Pathway Analysis

An examination of the metabolic and enzymatic pathways of *Halorhabdus utahensis* revealed further discrepancies within the annotation engines. The pentose phosphate pathway, a metabolic pathway that has variant forms in many other halophilic archaea [20], presents a clear case of this. The SEED database provided a KEGG pathway utilized in the analysis of this pathway; the annotation engine had marked various enzymes that were present in the organism, many of which are necessary for the functioning of the pathway (Figure 4). Four enzymes that catalyze the formation of key intermediaries were marked as absent.

However, further analysis demonstrated that one of these proteins, with E.C. number 5.3.1.6 and a predicted function as a ribose 5-phosphate isomerase, was indeed present in the genome. Using the student-authored programs, we determined that both the IMG and Manatee databases had predicted the enzyme's presence, although RAST had not. BLAST searches of these results provided confirmation that the amino acid sequences called by the annotation engines were indeed consistent with the sequence of a ribose 5-phosphate isomerase.

Both an E.C. number and text-based search of the gene calls of the databases for enzymes 1.1.1.49 (glucose 6-phosphate dehydrogenase) and 3.1.1.31 (6-phosphogluconolactonase) revealed that none of the annotation engines had predicted that their presence in the genome. Likewise, a BLAST comparison between the *H. utahensis* genome and protein sequences *Escherichia coli*, *Mycobacterium leprae*, and *Saccharomyces cerevisiae* yielded poor results (e-

values greater than .3 for both enzymes), indicating that neither of these proteins is likely to be found in the genome. However, similar methods for the enzyme 1.1.1.44 (phosphogluconate dehydrogenase) gave favorable e-values (8e-07 and 6e-06 for *E. coli* and *S. cerevisiae*, respectively), despite the fact that none of the databases had predicted that this enzyme was present in the genome. Although all three annotation engines support the conclusion that the pentose phosphate pathway is incomplete in *H. utahensis*, the databases disagree as to which genes are missing in the organism.

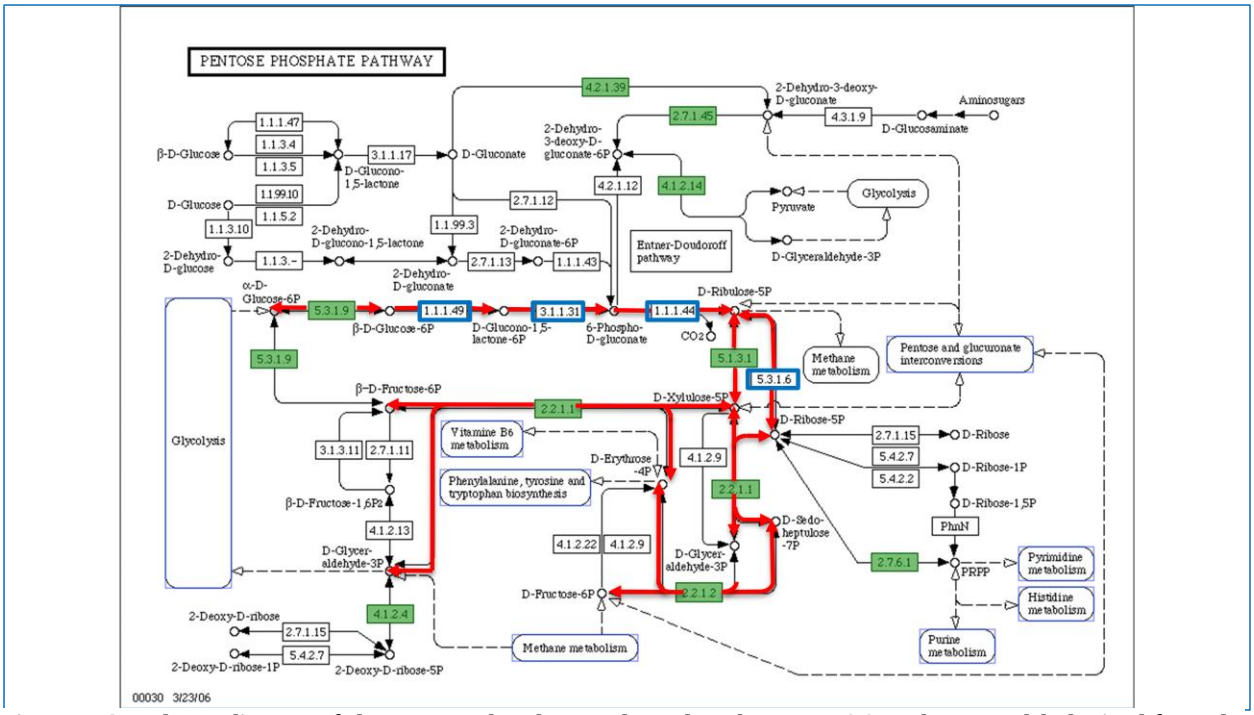

Figure 4. A pathway diagram of the pentose phosphate pathway based on a KEGG Pathway model obtained from the SEED database. The main oxidative and non-oxidative sections of the pathway are highlighted in red, while missing enzymes are highlighted in blue boxes and enzymes contained in the organism are marked green.

Discussion

The number of discrepancies between the genome annotations demands further scrutiny of the reliability of automated annotation. While it is very likely that the 1471 gene calls that are identical matches in all three databases reflect the correct annotation, the wide variation in gene calls reveals a general inconsistency in annotation strategy and software. The disagreements are particularly apparent in the variety of start codons of the gene calls. The annotations predicted the same stop codon but a different start codon in 1293 expected genes. Therefore, the algorithms and techniques each annotation service utilizes to determine start codons are sufficiently different to significantly influence the number of gene matches; in contrast, the tools each database uses to predict stop codons appear to be more consistent and accurate.

The variations between the different genome annotations reveal the limitations of automating genome analysis. The databases' annotations of the enzymatic pathways of *H. utahensis* particularly reveal the discrepancies between the different annotation engines. While all three annotation services contain the same nucleotide data, their interpretations of the data is very different, as evidenced by the annotation engines' conflicting conclusions about the presence and absence of the enzymes that form the pentose phosphate pathway. Not only did the

annotations disagree with regard to the presence of particular enzymes, but all three predicted that an enzyme (1.1.1.44) was not present when in fact a BLAST search indicated that there was likelihood that the gene for the protein is indeed located in the *H. utahensis* genome. Therefore, it is clear that all three automatic annotations are susceptible to error and inaccuracies; even comparisons between the different annotations will not provide a perfect solution to the problem.

This ambiguity and uncertainty could be largely rectified by manual, human curation of every genome; this time-consuming and cost-prohibitive process, however, is unrealistic. In order to ensure that all automated annotation software can provide a relatively accurate and trustworthy set of genome predictions, a set of standards should be established to ensure that every annotation meets a set of universal requirements. Standardization has become a necessity with the advent of "high throughput technologies" [8]. The astounding volume of information now available to scientists cannot possibly all be examined thoroughly by experts; automated annotation is a necessity, but even as such it must be treated as a potentially dangerous tool that can provide inaccurate, and therefore misleading, information. Quality control must be implemented to ensure that the inevitable inaccuracies are lessened and less likely to be harmful.

However, much of the techniques and approach that each annotation engine undertakes to analyze a genome will likely remain unique; in fact, until the algorithms and programs of automatic annotation software grow sophisticated to the point at which the number of errors they produce is statistically insignificant, multiple different databases should be considered when analyzing a genome in order to obtain the most precise and reliable annotation possible. Comparison of various databases remains necessary to identify and remedy errors that could spoil analysis.

Thus, the scientific community would be wise to heed its own wariness of trusting a single interpretation of an issue; the annotation of a genome by one annotation service is not sufficient to obtain a valuable analysis. Both a set of standards for all annotation systems and comparison of multiple annotations are necessary to best facilitate use of the incredibly powerful opportunities offered by the field of genomics.

References

1. Daniel H (2002) Genomics and proteomics: Importance for the future of nutrition research. British Journal of Nutrition 87: 305-311.

2. Verrips CT, Warmoeskerken MMCG, Post JA, et al. General introduction to the importance of genomics in food biotechnology and nutrition. Current Opinion in Biotechnology 12: 483-487.

3. McLaren JS (2000) The importance of genomics to the future of crop production. Pest Management Science 56: 573-579.

4. Eisen JA, Fraser CM (2003) Phylogenomics: Intersection of evolution and genomics. Science 300: 1706-1707.

5. Collins FS, Green ED, Guttmacher AE and Guyer MS (2003) A vision for the future of genomics research. Nature 422: 835-847.

6. Lopatto D, Alvarez C, Barnard D, Chandrasekaran C, Chung HM, et al. (2008) Genomics Education Partnership. Science. 322: 684-685.

7. Campbell AM, Ledbetter MLS, Hoopes LLM, Eckdahl TT, Heyer LJ, et al. (2007) Genome Consortium for Active Teaching: Meeting the Goals of BIO2010. CBE-Life Sci. Educ. 6: 109-118.

8. 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.

9. Brazma A (2001) On the importance of standardization in the life sciences. Bioinformatics 17: 113-114.

10. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, et al. (2006) The integrated microbial genomes (IMG) system. Nucleic Acids Research 34: 344-348.

11. Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, et al. (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Research 36: 528-533.

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

13. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research 33: 5691–5702.

14. Creasy T, Angiuoli S, Maharkar A, Felix V, Davidsen T, et al. (2007) J. Craig Venter Institute Manatee and the Annotation System Architecture. (http://manatee.sourceforge.net/pdf/Manatee_Presentation2005.pdf. Accessed 8 December 2008.

15. J. Craig Venter Institute (2008) JCVI Annotation Service. http://www.jcvi.org/cms/research/projects/annotation-service/overview/. Accessed 8 December 2008.

16. Max Win (2008) Available at: http://www.bio.davidson.edu/courses/genomics/2008/Win/ec/.

17. Will DeLoache (2008) Available at: http://gcat.davidson.edu/Wideloache/Webfiles/ecNum Blast.html.

18. Will DeLoache (2008) Available at: http://gcat.davidson.edu/Wideloache/Webfiles/AnnotationS earcher.html.

19. The wiki page can be accessed at: http://gcat.davidson.edu/GcatWiki/index.php/Halorhabdus_utahensis_Genome.

20. Falb M, Müller K, Königsmaier L, Oberwinkler T, Horn P, et al. (2008) Metabolism of halophilic archaea. Extremophiles 12: 177-196.

Acknowledgments

We thank Cheryl Kerfeld and Edwin Kim, our contacts at JGI; Matt DeJongh of Hope College, our contact for help with the SEED and RAST; and Ramana Madupu at the J. Craig Venter Institute for help with Manatee. We thank Jonathan Eisen (University of California Davis) and Gary Stormo (Washington University) for their advice and help with the project. We thank Kjeld Ingvorsen of the Biologisk Institut at Aarhus Universitet, Denmark, for verifying our hypotheses regarding the viability of enzymatic pathways. Finally, we also thank Chris Healey at Davidson College for both ordering and growing *H. utahensis* locally.