Comparison of JGI, RAST, and JCVI Automated Genome Annotation Tools through the Examination of the *Halorhabdus utahensis* Genome

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

Biology Department, Davidson College, Davidson, NC, USA
[1] Mathematics Department, Davidson College, Davidson, NC, USA
* To whom correspondence should be addressed

Abstract

The implementation of bioinformatics tools to produce automated annotations is a great advancement for the field of genomics, because it gives researchers more information to decode their genome and allows them to do so more rapidly. However, there are multiple automated annotation tools, and each completes the task in a slightly different manner. In this investigation, we compare the three annotations of the *Halorhabdus utahensis* genome produced by annotation systems supported by JGI, RAST, and JCVI. We found that these three annotations contained fundamental differences in number of genes predicted, gene length and start codon used. We also conducted closer analyses at certain predicted genes and biological pathways which yielded interesting results and conundrums that leave the possibility for further testing to improve these annotation systems.

Introduction

The field of genomics revolves around the sequencing and then subsequent annotating of the genetic information of organisms. In recent years, sequencing has become increasingly efficient and cost-effective. The development of methods such as those that allow for large-scale parallel sequencing by synthesis from amplified fragments of DNA have given researchers the ability to sequence up to 1 Gb in a single run [1]. Such advancements in sequencing technology have produced many sequenced genomes.

However, these sequences are untapped resources without the ability to decipher their meaning. The decoding or annotation of the genomes allows us to predict the location of protein-coding genes and therefore which proteins the organism is able to produce. Thus, annotation can also give us insight into the biological pathways that organisms use, new and undiscovered biological processes, and phylogenetic information. This information ultimately allows for a better understanding of how organisms are able to survive in their respective environments. Unfortunately, annotation of genetic data is not an easy task. Initially, labor and time-intensive manual annotation by wet-lab characterizations was the only viable way to annotate a genome. However, the advent of bioinformatics has allowed for faster annotation of genomes through automated annotation hybridized with manual annotation. That is, the implementation of computer science and mathematical tools to allow researchers to quickly compare non-annotated genomes to existing wet-lab characterizations. These tools extrapolate the data of wet-lab characterizations to show how well certain genes of the genome match the characterizations [2]. This gives researchers a preliminary idea of which genes are likely to exist in the genome they are annotating and allow them to more efficiently hand-curate the genome.

Three such automated annotation tools are JGI, RAST, and Manatee. The Integrated Microbial Genome (IMG) system powered by Joint Genome Institute (JGI) of the Department of Energy is an open resource for the annotation of all publicly available genomes. It predominantly uses NCBI's RefSeq as a source of "publicly available genomes" [3]. In its current version (2.6), IMG contains 4207 genomes for comparative analysis. Rapid Annotation using Subsystems Technology (RAST) of the Fellowship for Interpretation of Genomes (FIG) and other organizations is an automated annotation system that uses a database of FIG hand-curated "subsystems" and protein families (FIGfams) which are based on these subsystems for the computation [2]. Manatee of the J. Craig Venter Institute (JCVI) is an annotation tool for viewing and alteration of the initial automatic annotation for prokaryotic genomes. The annotations of Manatee are run through the JCVI Annotation Service, which uses the GLIMMER system (a powerful tool that was found to identify 97-98% of all genes when compared with published

annotation) as the major tool to identify genes [4]. Then JCVI runs a search of these protein predictions against proteins from other protein databases.

This investigation compared the automated annotation of *Halorhabdus utahensis AX-2* produced by these three tools. *H. utahensis* is a gram-negative halophile archaeon that lives in the Great Salt Lake in Utah. *H. utahensis* is an extremophile that grows optimally at 27% (w/v) NaCl; at the time of discovery, *H. utahensis* had the highest reported percent salinity growth optimum of any organism [5]. The ability of *H. utahensis* to live in such a hazardous environment leads one to believe that its genome is a potential source of information to discover and understand unique proteins and biological processes that were engineered by natural selection [6]. Therefore, it is imperative to have an annotation that is accurate and effective. An efficient automated annotation system is a key component to such an annotation; therefore, this study focuses on the annotation output by three major automated annotation tools: JGI, RAST, and Manatee.

Materials and Methods

JGI sequenced the genome of *Halorhabdus utahensis* by the Sanger method and then ran the sequence through JGI's IMG automated annotation tool before we received the genetic data. We accessed the annotated data through JGI's website (http://img.jgi.doe.gov). We then began hand-annotation of RNA genes and then randomly selected genes (with and without predicted function as called by JGI) by the guidelines of the JGI's educational annotation handbook (which can be found at http://gcat.davidson.edu/GcatWiki/index.php/Gene_Annotation_Template). Aside from JGI, other online Databases and tools including BLAST, CDD, TIGRfam, TMHMM, SignalP, PSORT, Phobius, Pfam, PDB, T-coffee, KEGG and ExPASy were used to find and verify data concerning COGs, predicted protein location, signal peptide probability, phylogenetics, biological pathways and EC numbers.

We then sent the *H. utahensis* genome to RAST (http://rast.nmpdr.org/) and Manatee (http://www.tigr.org/tigr-scripts/prok_manatee/shared/login.cgi). Upon receiving their annotations, which we accessed through their websites, we compared basic annotation results (i.e. number of genes found). Next, we created computer

programs to compare the genes that were found by each annotation system. As a consequence, we developed a computer program to compare the start and stop sites of the three annotation tools. We examined a number of genes that did not match across all three annotation tools by hand in order to find fundamental reasons behind the mismatches. In these annotations we used NCBI's BLAST tools (i.e. BLASTp, BLAST2) to find protein matches and alignments of the information provided by the three annotations.

We hand-annotated various biological pathways to see which pathways were present or non-existent in *H. utahensis* and how useful the annotations were in doing so. We viewed the pathways through the KEGG pathway viewer and through the RAST-customized KEGG metabolic analysis tool. Presence of enzymes was determined by using a combination of tools. We used ExPASy to retrieve EC number and enzyme names. We then used group-developed computer programs. We developed one program so that it searched for common amino acid FASTA sequences from halophilic organisms by EC number and then BLASTed them against *H. utahensis*'s genome. We developed another so that it searched for the presence of a various EC number in any of the three automated annotation. We developed a final program to search by protein name to see if the searched name was present in any of the annotations.

We chronicled our progress online through a WikiMedia powered site, which can be accessed at:
http://gcat.davidson.edu/GcatWiki/index.php/Halorhabdus_utahensis_Genome.

Results

Three-way Annotation Comparison

In our initial group overview and analysis of the three annotation systems, we recognized several fundamental differences in their respective annotations. First, we noticed a significant difference in the number of genes that were called by each annotation. Manatee predicted the most open reading frames (ORFs) with 3254, while JGI and RAST called 3097 and 2898, respectively. After noticing this, we looked for reasons of these discrepancies. We found that there were only 1471 exact gene prediction

matches (see Figure 1); that is, cases in which all annotations called the same open reading frame (or the same locations for the start and stop codons). However, we also found that 2764 stop codon locations were shared by all three annotations. Therefore, there were 1293 instances in which all three annotations had the same stop codon, but at least one annotation had a different start codon location. In total, Manatee had the greatest number of unique stop codons with 254 (see Figure 2).

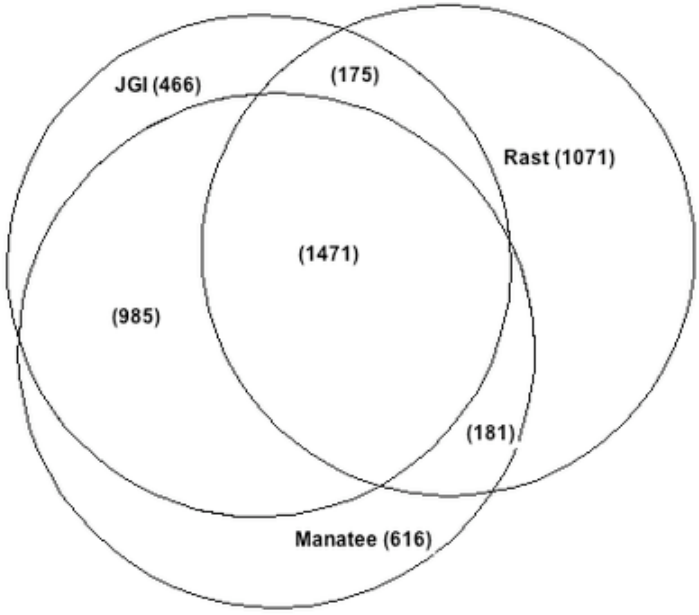


Figure 1: Venn diagram displaying the exact gene matches (start and stop codons) shared between the three annotation tools, matches between only two, and independent gene calls.

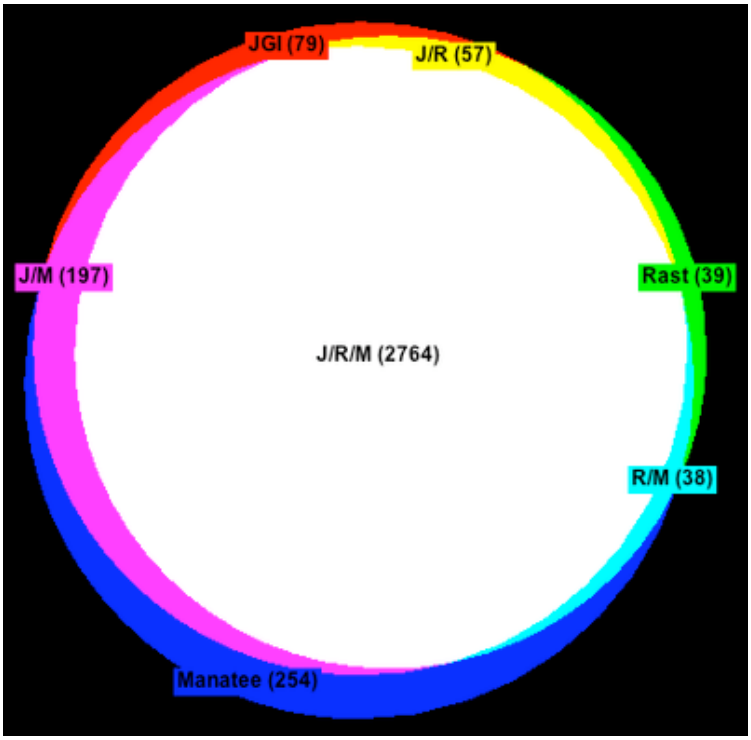


Figure 2: Venn diagram displaying the stop-codon matches between all three annotation tools, matches between only two, and totally unique stop-codon annotations.

We then compiled a list which included gene annotations where either two of the systems called a gene identically and the other one missed it completely, or in which the other system called the same stop codon, but had a different start codon. We then examined a number of these discrepancies manually and found a number of patterns that helped to clarify the difference in gene predictions. We noticed that when we looked at these annotation discrepancies by hand that RAST very often used alternative start codons, whereas JGI and Manatee had greater propensities to use the common ATG start codon. We found that RAST used alternative start codons in 39.0% of its gene predictions, whereas JGI and Manatee used 14.3% and 19.9%, respectively (see Table 1). TTG and GTG were the only alternative start codons used; CTG was not used.

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%

Table 1: Table displaying the start codons used by each annotation tool, showing the comparison of the use of the usual ATG start codon and alternative start codons.

We noticed that in the cases when RAST chose to use an alternative start codon rather than ATG, the alternative was almost exclusively upstream from the ATG start, resulting in a longer ORF. So, we then compared the predicted gene lengths of the

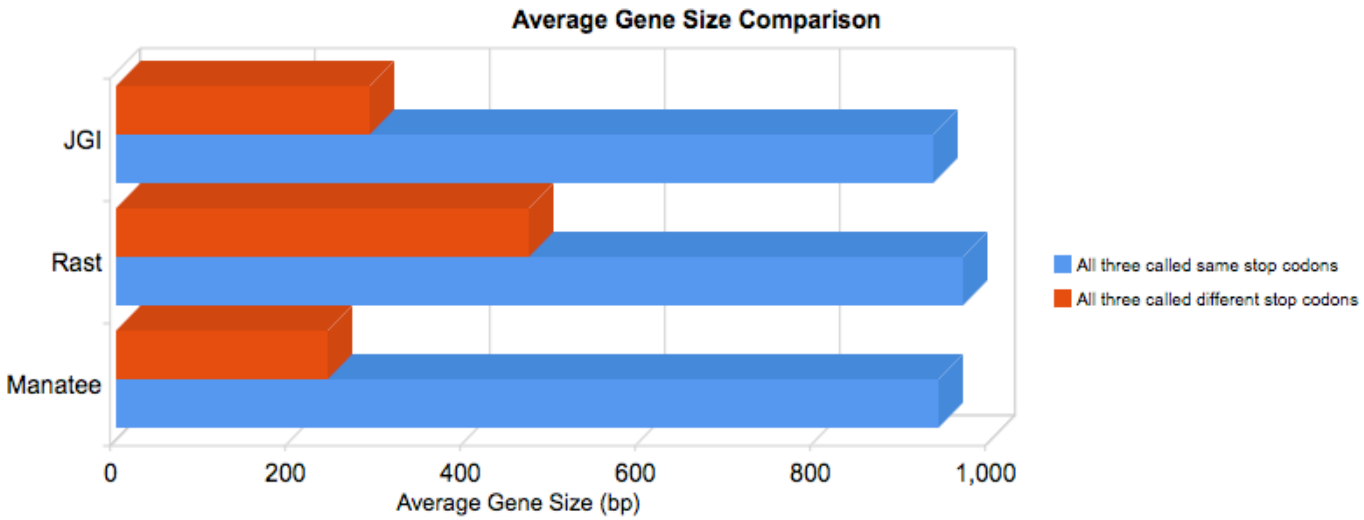


Figure 3. Graph that compares average gene length of the three annotation sites for when each shared stop codon locations (blue) and when each called a different stop codon (red)

predictions in which all three annotations shared the same stop codon location. We found that JGI had the shortest genes on average with a 934 base pair average, Manatee had the second most with 940 base pairs, and RAST the longest with 967. We also looked at the instances when the annotations called a gene that the others did not. We found that the average gene lengths were significantly shorter, while RAST's were still longest: JGI with 290 base pairs, Manatee with 242 base pairs and RAST with 472 base pairs. Overall, Manatee had the shortest average gene length with 844.9, then JGI with 869.9 and RAST with 941.8 (see Figure 4).

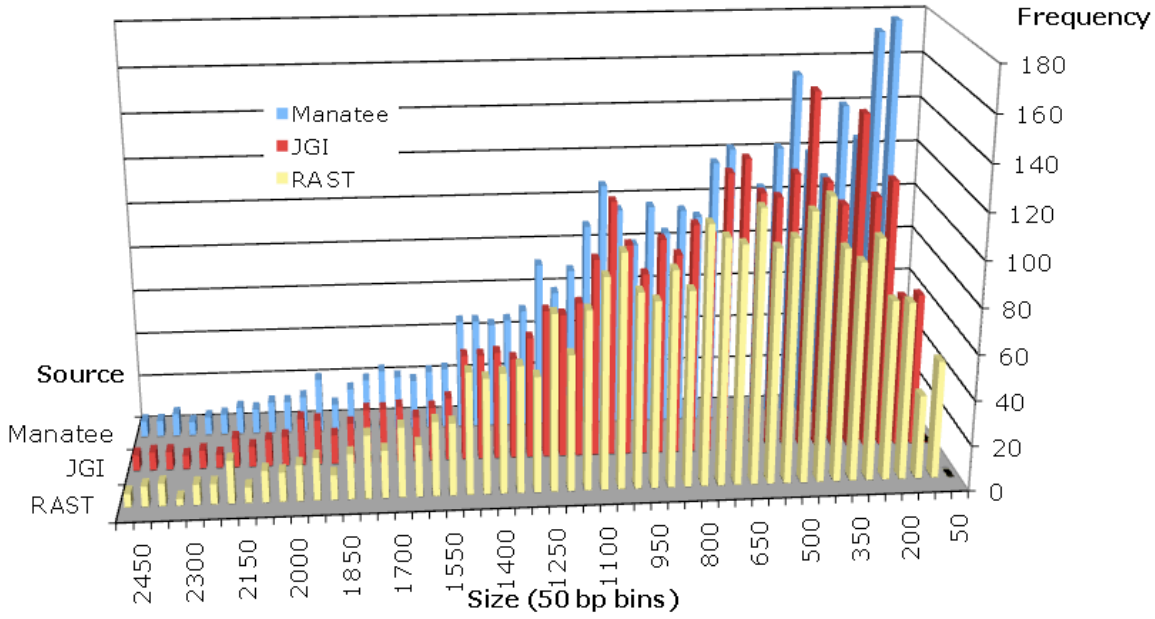


Figure 4: Graph displaying the comparison of the frequency of various gene sizes in base pairs in the three automated annotations

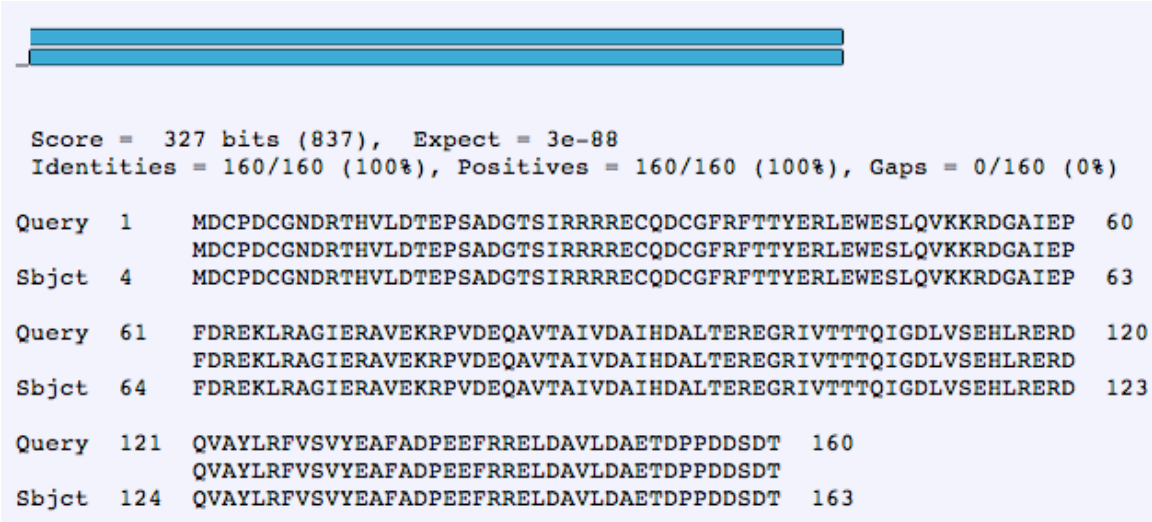
Detailed Gene Prediction Analysis

While doing the overview analysis of these sites and doing manual annotations, we came to find various gene predictions that stood out for various reasons. One such gene prediction was one that called for the transcriptional regulator nrdR. We found this gene originally as a call by Manatee at base pairs 3109722…3110204 (+); the other two

annotations called nothing in this region. We used BLASTp to verify Manatee's prediction and found it to be legitimate call, with an expected value of 2e-56. (see Figure 5). The match from BLASTp was also from another halophile, *Natronomonas pharaonis*, strengthening the validity of Manatee's prediction. Being a transcriptional regulator, we expected that the other annotations should have predicted the gene as well, so we searched for it in different locations within the other annotations. We found that the other two annotations did indeed have a prediction for nrdR with a good expected value, but in a completely different location. RAST called the same gene at 7274..7765 (+) and JGI did at 7283...7765 (+); it important to not that they called the same stop codon, but different slightly on the start codon (RAST used the alternative start codon GTG). The very close similarity of these two predictions was verified by a BLAST2 alignment, which showed that there was only a 3 amino acid discrepancy (caused by the difference in start codons) (see Figure 6).

```
>ref|YP_327708.1| transcriptional regulator NrdR [Natronomonas pharaonis DSM 2160]
 sp|Q3IM07|NRDR_NATPD Transcriptional repressor nrdR
 emb|CAI50861.1| conserved hypothetical protein [Natronomonas pharaonis DSM 2160]
Length=162

 GENE ID: 3694633 NP6162A | transcriptional regulator NrdR
[Natronomonas pharaonis DSM 2160] (10 or fewer PubMed links)

 Score =  220 bits (561),  Expect = 2e-56, Method: Compositional matrix adjust.
 Identities = 106/160 (66%), Positives = 130/160 (81%), Gaps = 0/160 (0%)

Query  1    MDCPDCGNDRTHVLDTEPSADGTSIRRRRECQDCGFRFTTYERLEWESLQVKKRDGAIEP  60
            M+CPDCGN RT V+DT  S+DG S+RRRRECQ C FRFTTYER EW+SLQVKKRDG IE
Sbjct  1    MNCPDCGNGRTRVIDTGASSDGASVRRRRECQRCSFRFTTYERPEWKSLQVKKRDGTIES  60

Query  61   FDREKLRAGIERAVEKRPVDEQAVTAIVDAIHDALTEREGRIVTTTQIGDLVSEHLRERD  120
            FD++KLR GIERAVEKR V E  VTA+VD I   L +RE RIV+++ IG+LVSE+LR  D
Sbjct  61   FDQQKLRTGIERAVEKRGVAETTVTALVDDIESELQDREARIVSSSLIGELVSENLRTLD  120

Query  121  QVAYLRFVSVYEAFADPEEFRRELDAVLDAETDPPDDSDT  160
            +VAY+RFVSVY+AF++P+EF +ELDAVL AE D  + S++
Sbjct  121  KVAYIRFVSVYKAFSEPQEFLKELDAVLGAELDDFEASNS  160
```

Figure 5. BLASTp result used to verify the gene prediction by Manatee of nrdR



```
 Score =  327 bits (837),  Expect = 3e-88
 Identities = 160/160 (100%), Positives = 160/160 (100%), Gaps = 0/160 (0%)

Query  1    MDCPDCGNDRTHVLDTEPSADGTSIRRRRECQDCGFRFTTYERLEWESLQVKKRDGAIEP  60
            MDCPDCGNDRTHVLDTEPSADGTSIRRRRECQDCGFRFTTYERLEWESLQVKKRDGAIEP
Sbjct  4    MDCPDCGNDRTHVLDTEPSADGTSIRRRRECQDCGFRFTTYERLEWESLQVKKRDGAIEP  63

Query  61   FDREKLRAGIERAVEKRPVDEQAVTAIVDAIHDALTEREGRIVTTTQIGDLVSEHLRERD  120
            FDREKLRAGIERAVEKRPVDEQAVTAIVDAIHDALTEREGRIVTTTQIGDLVSEHLRERD
Sbjct  64   FDREKLRAGIERAVEKRPVDEQAVTAIVDAIHDALTEREGRIVTTTQIGDLVSEHLRERD  123

Query  121  QVAYLRFVSVYEAFADPEEFRRELDAVLDAETDPPDDSDT  160
            QVAYLRFVSVYEAFADPEEFRRELDAVLDAETDPPDDSDT
Sbjct  124  QVAYLRFVSVYEAFADPEEFRRELDAVLDAETDPPDDSDT  163
```

Figure 6. BLAST2 alignment of RAST and JGI nrdR gene predictions

Finally, we checked to see if the new location was also called by Manatee; it was not. There were also no other genes called in that region. Therefore, there were two different locations, and both were never called by a single annotation.

Biological Pathway Analysis

After doing analyses of single genes, we widened our scope and began to annotate biological pathways. In one such analysis, we looked at the metabolism of chitin. We found a potential coding region for chitinase, and knew that our organism had a source of chitin from brine shrimp that inhabit the Great Salt Lake. As a result, we had done a preliminary test to see if *H. utahensis* could grow on N-acetyl-D-glucosamine (NAG), the monomer of chitin, but *H. utahensis* could not. Therefore, we chose to take a closer look at this pathway, because it was odd that the organism would code for chitinase, yet not be able to grow on its respective monomers.

The pathway (see Figure 7) was adapted to *H. utahensis* by looking for the existence of each enzyme in the pathway in any of the annotations or by hand-curation using our computer program (see materials and methods). We found that in addition to having chitinase to break down chitin to NAG, *H. utahensis* also had the enzyme Chitin deacetylase to transform chitin into another polymer, chitosan with a waste product of

acetate. From NAG, the organism is also predicted to produce UDP-N-acetyl-D-mannosaminouronate as UDP-N-acetyl-D-galactosaminuronate, as well as to reproduce chitin.

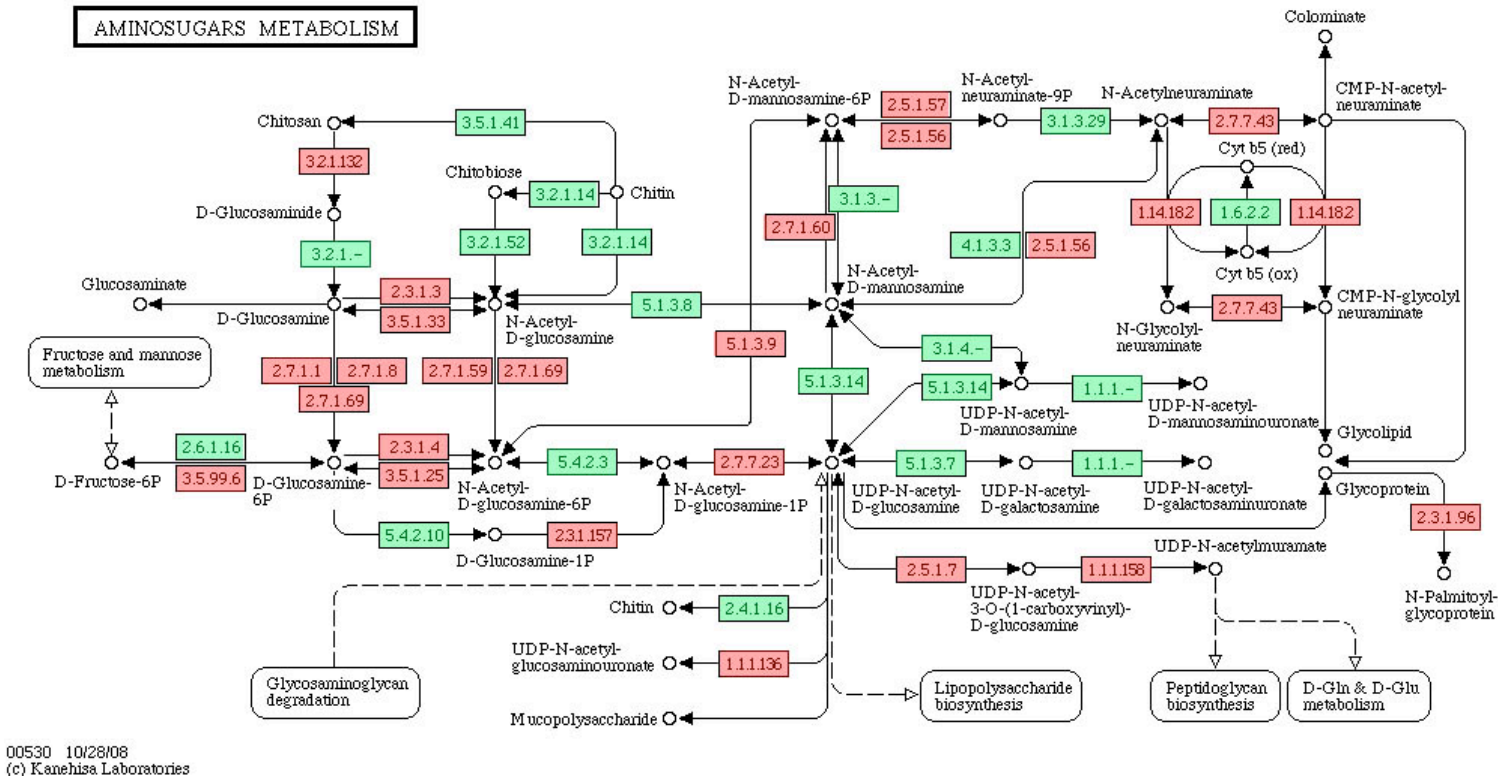


Figure 7: A modified pathway of the aminosugars pathway from KEGG that displays the metabolism of Chitin in *H. utahensis*. Green denotes an enzyme that was found in one of the annotations of by hand-curation; red denotes an enzyme that was found not to exist

Discussion

The results of this investigation give an initial insight into the differences between these three systems and also where the systems could be improved to provide a better automated annotation service. These insights call for increased side-by-side research between the algorithms that each annotation system uses, the predictions the systems yield and physiological, wet-lab testing to produce in order to produce data that would improve these three systems. Finally, this investigation also gave us some preliminary information about the biological workings of *H. Utahensis*.

The analysis of the nrdR gene prediction helps to highlight some of the discrepancies in the three annotations. It shows that these annotation systems are imperfect because they miss perfectly valid gene predictions, which is most likely due to the difference in characterizations that they use to compare the genome with. However, by synthesizing the three annotations and analyzing them manually, one is able to find more possible valid gene locations. Therefore, use of a corporate annotation coupled with manual annotation may be helpful to genomic researchers in the short-term. Analysis of the nrdR gene also gave an example of RAST's propensity to use alternative start codons. In this case, the use of the alternative start codon and therefore the addition of 3 amino acids did not change the gene prediction. However, the use of alternative start codons and the subsequent lengthening of the ORF would invariably lead to changes in gene predictions. It would be interesting to look further into the use of alternative start codons with physiological testing to see how the alternative start codons affected the accuracy of the annotation.

The analysis of biological pathways in *H. Utahensis* like the metabolism of chitin in the aminosugars pathway provided information about the organism's ability to perform certain biological processes and the ability of these systems to annotate pathways. Analysis of the aminosugars pathway showed that chitin could indeed be metabolized into NAG, its monomeric component and another polymer, chitosan. From NAG the final annotation showed that *H. Utahensis* could form UDP-N-acetyl-D-glucosamine, which could be used to reform chitin or be used in the biosynthesis of lipopolysaccarides. Therefore, *H. Utahensis*, a gram-negative archaeon, may use either of these structurally important compounds; however, it is more likely that chitin is metabolized to be used in the production of lipopolysaccarides, which are major components of the outer membranes of gram-negative bacteria (7). This is an instance where further physiological testing would help to better clarify how the organism utilizes chitin.

The process of pathway annotation through the use of the automated annotations proved to be somewhat difficult. Only RAST provides a KEGG component of its annotation which in theory highlights all enzymes called by RAST that had an EC number in the annotation. However, we found that this component often did not highlight its own calls. It would be helpful if this problem was solved, and if there was an option to

edit pathways by hand. In addition, the annotations sometimes did not label their calls with an EC number, which made it more difficult to find out whether an enzyme existed within an annotation. In addition, we found in many cases that our computer program found perfectly valid BLAST matches that supported the existence of enzymes that were overlooked by all three of the annotation systems. These cases further support the need for manual annotation as oversight for the errors that the automated annotations inevitably make.

By and large, this investigation yielded interesting findings about the tools provided by automated annotation and provided a good first look at the *H. Utahensis* genome. These three annotation systems are fundamentally different and produced three reasonably different annotations. We found that synergizing the three automated annotations and using manual annotation as oversight seemed to produce the most logical and viable annotations. In the future, however, it is important that physiological testing is performed to pinpoint weaknesses of each system for improvement. An expansive investigation that looked at each system and then developed a new system that combined strengths of each into one program may be the most effective means to an improved system. In any case, this investigation has provided insight into the ability and further potential of automated annotation to greatly speed up research in the field of genomics.

Acknowledgements

Thanks to Cheryl Kerfeld and Edwin Kim of JGI for their help and support concerning JGI, Jonathan Eisen of the University of California at Davis for his insight concerning the project, Gary Stormo of Washington University at St. Louis for coming to Davidson and talking with us about Genomics, Matt DeJongh of Hope College for guidance concerning SEED/RAST, and Ramana Madupu of the J. Craig Venter Institute for help with Manatee. We also thank Kjeld Ingvorsen of The Institute of Biological Science at the University of Aarhus in Denmark for his help with physiological tests on *H. utahensis*. Finally, we thank Chris Healey at Davidson College for ordering and growing *H. utahensis.*

References

1. Hall, H. (2007) Advanced sequencing technologies and their wider impact in microbiology. The Journal of Experimental Biology 209: 1518-1525.

2. Overbeek, R.A., Aziz, R. K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O. (2008) The RAST server: rapid annotations using subsystems technology. BioMed Central Genomics 9: 1-15. <www.biomedcentral.com/1471-2164/9/75>.

3. Kyripudes, N.C., Markowitz, V.M., Szeto E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I.A., Dubchak, I., Anderson, A., Lykidis, A., Mavromatis, K., Ivanova N.N. (2007) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Research Advance Access*:* 1-6.

4. Delcher, A.L., Harmon, D., Kasif, S., White, O., Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Research 27: 4636-4641.

5. Ingvorsen, K., Tindall, B., Wainø, M. (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.

6. Dennis, P. P., Shimmin, L. C. (1997) Evolutionary Divergence and Salinity-Mediated Selection in Halophilic Archaea. Microbiology and Molecular Biology Reviews 61: 90-104.

7. Riggig, M.G., Kaufmann, A., Robins, A., Shaw B., Sprenger, H., Gemsa, D., Foulongne, V., Rouot, B., Dornand, J. (2003) Smooth and rough lipopolysaccharide phenotypes of Brucella induce different intracellular trafficking and cytokine/chemokine release in human monocytes. Journal of Leukocyte Biology 74: 1045-1055.

8. Hingamp, P., Brochier, C., Talla, E., Gautheret, D., Thieffry, D., Herrmann, C. (2008) Metagenome Annotation Using a Distributed Grid of Undergraduate Students. PLoS Biology 6(11): e296 doi:10.1371/journal.pbio.0060296.

Appendix

It is of important note that this research was conducted through the structure of an undergraduate biology course at Davidson College. There has been an increasing interest in involving undergraduates in genomic research, because of the opportunity for those with relatively limited biological background (i.e. undergraduates) to make meaningful impact in annotation of genomes with the aid of automated annotation (8). This research stands as testament to the work that undergraduates are capable of doing.