A tail of three annotation tools

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

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

Abstract

In this study, we compared the results of three gene annotation tools to better understand the *Halorhabdus utahensis* genome and the accuracy and reliability of these tools. We submitted the *Halorhabdus utahensis* genome to The Doe Joint Genome Institute (JGI), Manatee, and Rapid Annotation using Subsystem Technology (RAST) for gene annotations and analyzed their predicted ORFs. Our results indicated that all three tools called the same stop codon but different start codons for 1293 ORFs. In addition, Rast found fewer genes but had longer average gene length compared to JGI and Manatee suggesting that Rast's annotation tool might have a higher cut-off value for gene size. Furthermore, we found the Shine Dalgarno sequence and hand-curated several genes and metabolic pathways to validate some of the annotations.

Introduction

Genome sequencing has become faster and cheaper in recent years. As a result, many gene-prediction tools had been developed for annotating genomes. The accuracy and reliability of these tools varies. Each of these tools has its own strength and weakness. In this study, we are focusing on the *Halorhabdus utahensis's genome. Halorhabdus utahensis is* an aerobic halophile isolated from Great Salt Lake in Utah. B*y* comparing the annotated results of JGI, Manatee and RAST, we understand more about the *Halorhabdus utahensis's genome.*

Materials and Methods

Comparison of JGI, Manatee and Rast annotation tools

We obtained ORFs data from JGI, Manatee and Rast websites after their analysis was done. There were one large contig and four small contigs. We only compared ORFs from the large contig since only a number of genes were found on the small contigs.

For gene comparison, we developed various programs to compare the start and stop codons, compute average gene size, and find alternative start codons.

Hand Curation tools

For hand curation, we developed a text-based seach tool that returns protein function and sequences, and an EC number search tool that enables us to query an EC number, blast all the proteins associated with this EC number with the *H. utahensis* genome.

In addition, we also used a wide variety of web tool such as NCBI, BLAST, KEGG, CDD and Pfam.

Results

Comparison of ORFs across JGI, Rast and Manatee annotations

Annotation Engines	Total number of ORFs found
Manatee	3253

JGI	3097
Rast	2898

Figure 1. Total number of ORFs found by Manatee, JGI or Rast

Figure 1 shows that Manatee found about 200 more ORFs compared to JGI annotation engine and 355 more ORFs compared to Rast annotation engine. The differences among the three annotation engines were greater than we expected. To understand more about these differences, we compared ORFs found by the three annotations.

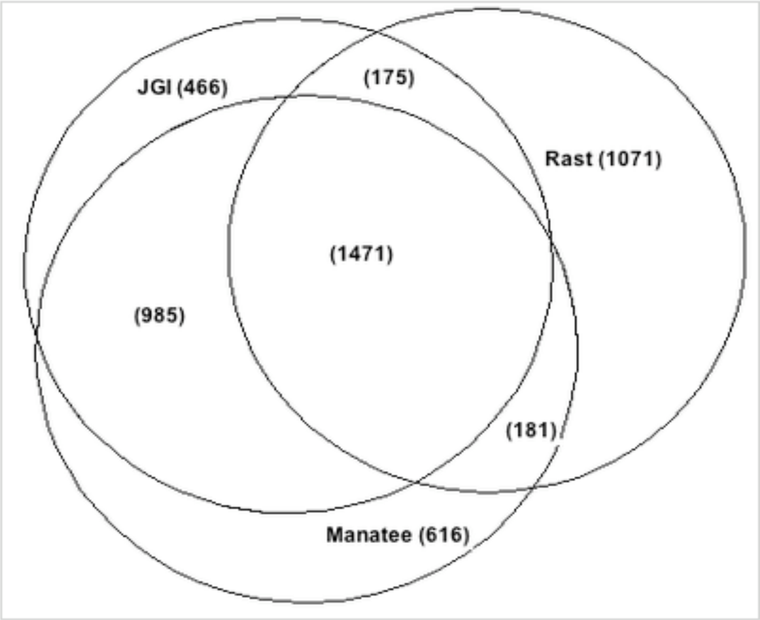


Figure 2. Venn diagram showing the number of ORF matches across the 3 annotations. Regions that overlap denote that the overlapping annotations found the same ORFs.

Our result, as shown in Figure 2, indicated that only 1471 ORFs were identical among JGI, Manatee and Rast annotations. From the Venn diagram in Figure 2, we observed that Manatee and JGI shared a large number of ORFs suggesting that of the three annotations, the annotation algorithms of JGI and Manatee were more similar to each other.

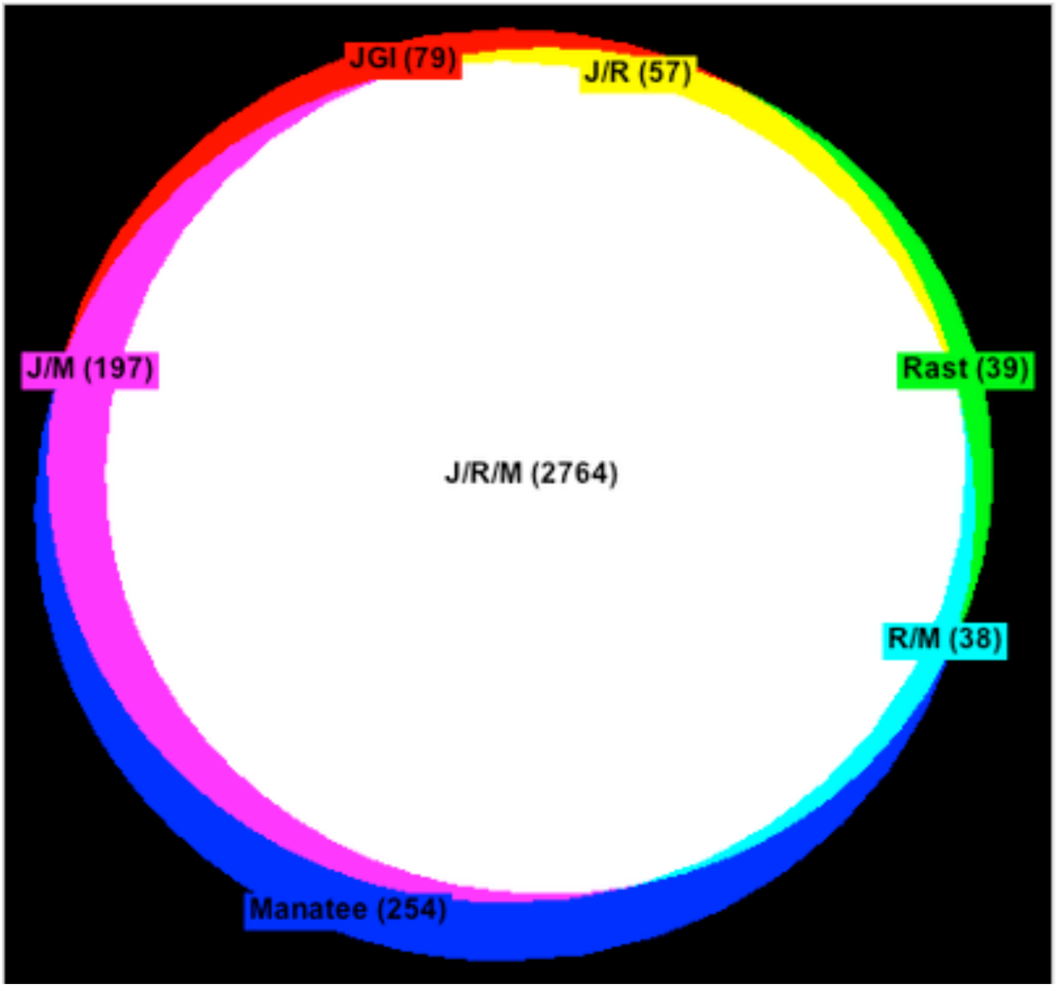
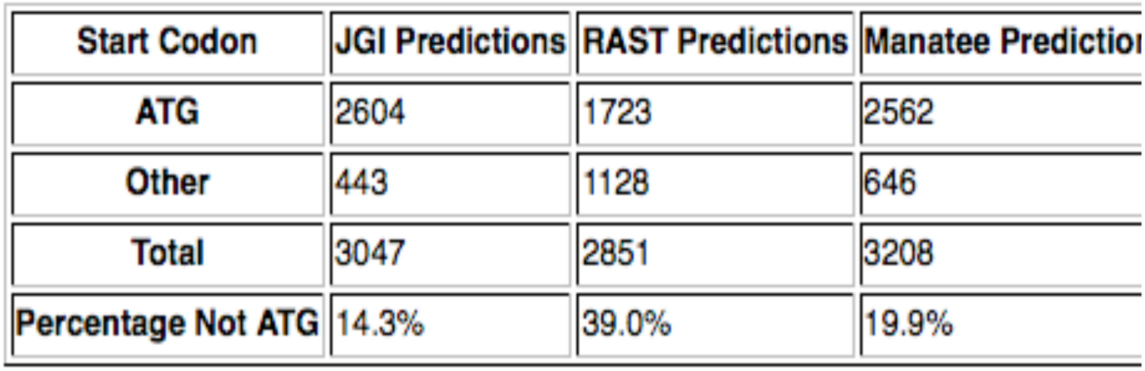


Figure 3. Venn diagrams showing the number of stop-index matches across the three annotations. Regions that overlap denote that the overlapping annotations called the same stop index and strand for a given gene.

When we only compared the stop index of the ORFs from the three annotations, we found that the overlapping region of JGI, Manatee and Rast significantly increased as shown in Figure 3. This result suggested that the three annotations agreed more on stop index than on start index.

In particular, when all three annotations called the same stop index, it appeared that Rast often called a different start index from that of JGI or Manatee. Figure 4 shows that Rast was about two times more likely to call alternative start codons compared to JGI and Rast.



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

Figure 4. Alternative start codon was either TTG or GTG. None was CTG.

In addition, the average gene length of JGI ORFs was 869.9. The average gene length of Rast ORFs was 941.8 and the average gene length of Manatee ORFs was 844.9. Rast and JGI found fewer ORFs than Manatee and the average gene size of Rast and JGI are greater than Manatee. These data suggested the reason why Manatee called more ORFs is because Manatee has a lower cut-off value for the size of the ORFs. Rast, in particular, has a much higher cut-off value for the size of the ORFs. As a result, Rast predicted fewer ORFs.

When all three annotations found the same stop codons but different start codons, do alternative start codons significantly affect the size of the gene? To answer this question, we compared the average gene length of the ORFs that have the same stop codon across the three annotations. We also looked at the average gene length of genes that were uniquely called by JGi, the average gene length of genes that were uniquely called by Rast and the average gene length of genes that were uniquely called by Manatee. The result was summarized in the following figure.

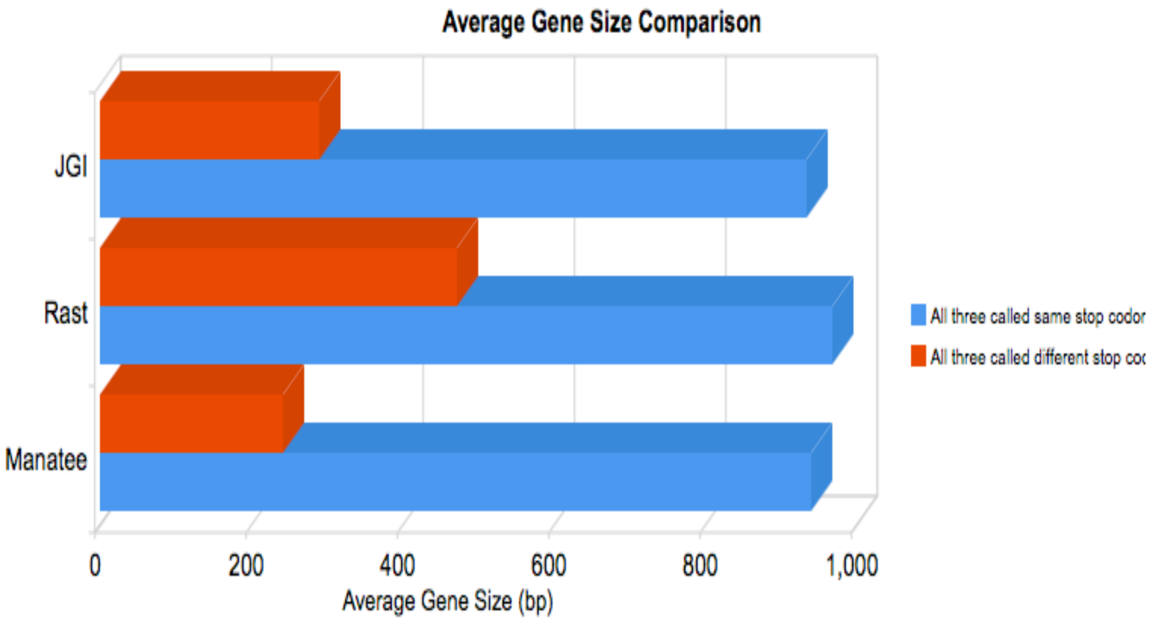
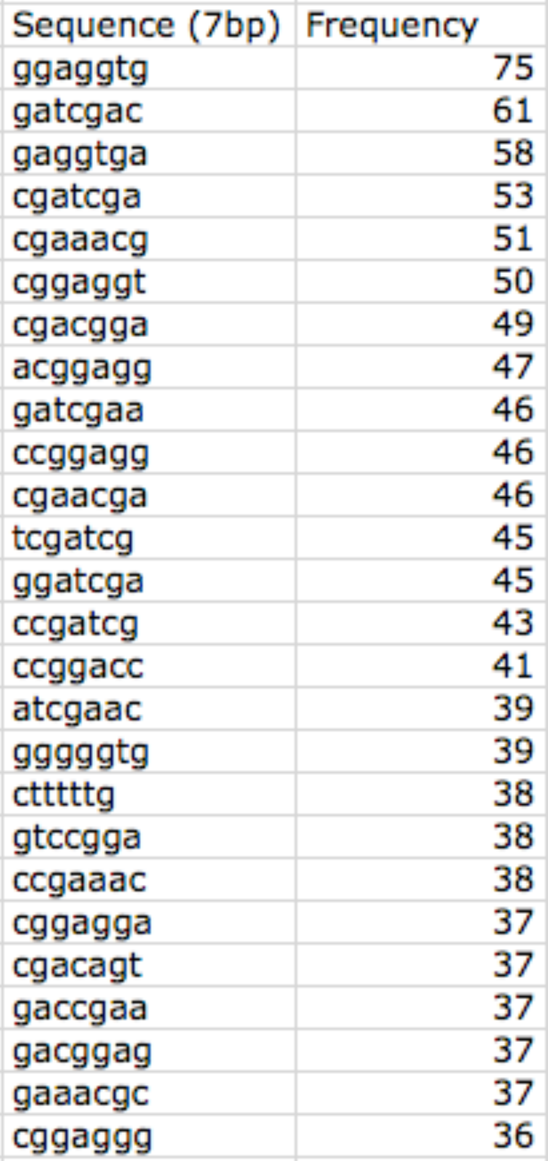


Figure 5. The blue bars denote the comparison of the average gene length of ORFs with the same stop codon across all three annotations. The red bars denote the comparison of the average gene length of ORFs uniquely called by JGI, Rast or Manatee.

Figure 5 shows that the average gene length of ORFs with the same stops condon across all three annotations was about the same. These data indicated that when all three annotations called the same stop codon but different start codons, the location of these start codons were close to each other. On the other hand, when we compared ORFs uniquely called by JGI, Rast or Manatee as indicated by red bars in Figure 5, we found that the average gene size of Rast was significantly greater than that of JGI and Manatee. This result again indicated that Rast did not pick up smaller ORFs by setting a higher cut off value for gene size.

Shina Dalgarno Sequence

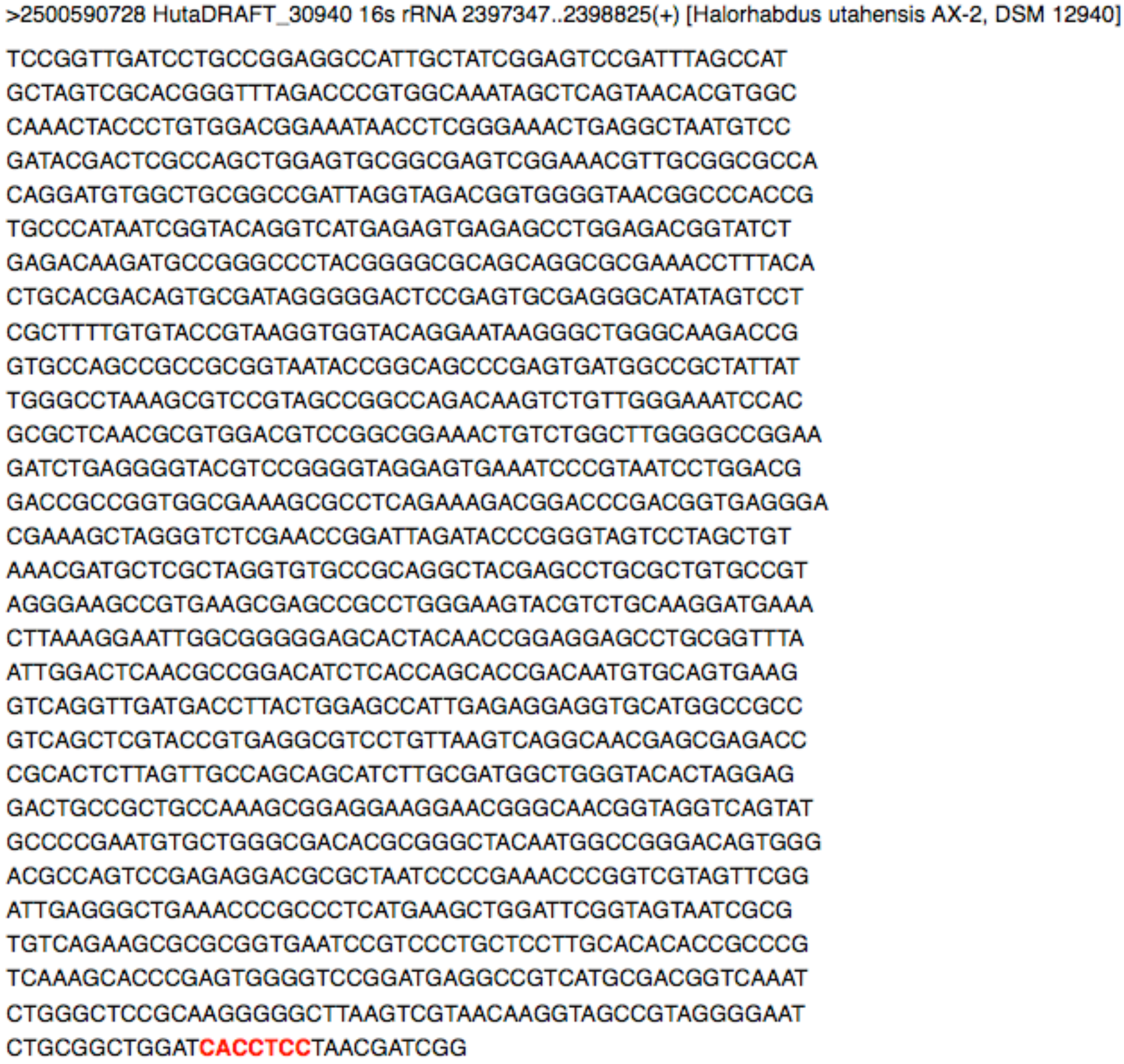
In order to find Shina Dalgarno Sequence, we need to find ribosome binding sites located upstream of the genes. We knew that RBS should be about 7bp long and should be relatively conserved. Our first step was to find the most conserved 7bp sequence 50 bp upstream of all the genes.



Sequence (7bp)	Frequency
ggaggtg	75
gatcgac	61
gaggtga	58
cgatcga	53
cgaaacg	51
cggaggt	50
cgacgga	49
acggagg	47
gatcgaa	46
ccggagg	46
cgaacga	46
tcgatcg	45
ggatcga	45
ccgatcg	43
ccggacc	41
atcgaac	39
gggggtg	39
ctttttg	38
gtccgga	38
ccgaaac	38
cggagga	37
cgacagt	37
gaccgaa	37
gacggag	37
gaaacgc	37
cggaggg	36

Figure 6. Conserved 7bp sequences located at 50 bp upstream of all the genes in JGI annotation.

The most conserved 7bp sequence located at 50bp upstream of all the genes in JGI, as shown in Figure 6, was GGAGGTG. GGAGGTG was also the most frequently occurred 7bp sequence in Manatee and Rast (data not shown). To verify that GGAGGTG is indeed the anti Shina Dalgarno sequence, we looked the DNA sequences of 16s rRNA and found the complementary Shina Dalgarno sequence CACCTCC.

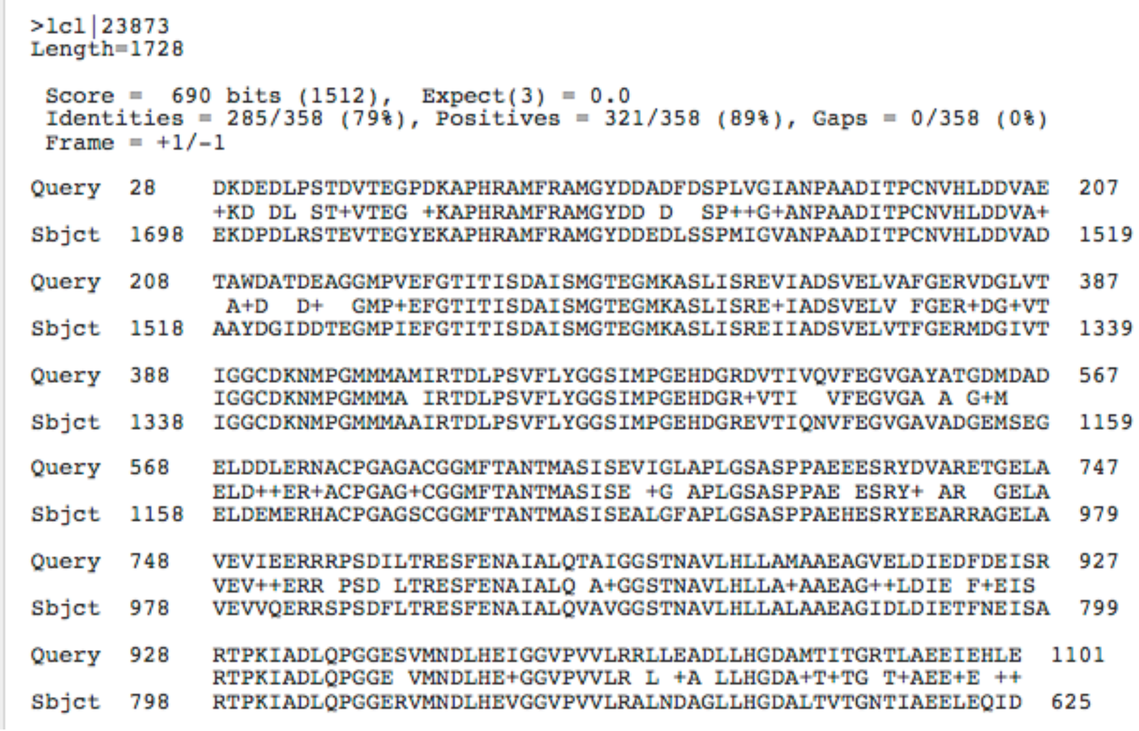


```
>2500590728 HutaDRAFT_30940 16s rRNA 2397347..2398825(+) [Halorhabdus utahensis AX-2, DSM 12940]
TCCGGTTGATCCTGCCGGAGGCCATTGCTATCGGAGTCCGATTTAGCCAT
GCTAGTCGCACGGGTTTAGACCCGTGGCAAATAGCTCAGTAACACGTGGC
CAAACTACCCTGTGGACGGAAATAACCTCGGGAAACTGAGGCTAATGTCC
GATACGACTCGCCAGCTGGAGTGCGGCGAGTCGGAAACGTTGCGGCGCCA
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
CTGCGGCTGGATCACCTCCTAACGATCGG
```

Figure 7. Nucleotide sequence of 16s rRNA with Shina Dalgarno sequence colored in red

My favorite Gene

One particular gene that interested me was dihydroxy-acid dehyratase with EC number 4.2.1.9. This gene was found on the negative strand of the main contig and started at nucleotide position 1849. One thing that was odd about this gene was that Manatee and Rast ended this gene at position 2 whereas Manatee ended the gene at position 1. Mostly importantly, there was no stop codon at the end of the ORF. Since the *H. utahensis* genome has 4 other smaller contigs, it is possible that the 3' end of this gene was located on one of these small scaldfolds. Thus, I blasted this gene against all microbes and the best match was the dihydroxy-acid dehyratase of the *H. utahensis* genome's close relative Haloarcula marismortui.



```
>lcl|23873
Length=1728

 Score =  690 bits (1512),  Expect(3) = 0.0
 Identities = 285/358 (79%), Positives = 321/358 (89%), Gaps = 0/358 (0%)
 Frame = +1/-1

Query  28    DKDEDLPSTDVTEGPDKAPHRAMFRAMGYDDADFDSPLVGIANPAADITPCNVHLDDVAE  207
             +KD DL ST+VTEG +KAPHRAMFRAMGYDD D  SP++G+ANPAADITPCNVHLDDVA+
Sbjct  1698  EKDPDLRSTEVTEGYEKAPHRAMFRAMGYDDEDLSSPMIGVANPAADITPCNVHLDDVAD  1519

Query  208   TAWDATDEAGGMPVEFGTITISDAISMGTEGMKASLISREVIADSVELVAFGERVDGLVT  387
              A+D  D+  GMP+EFGTITISDAISMGTEGMKASLISRE+IADSVELV FGER+DG+VT
Sbjct  1518  AAYDGIDDTEGMPIEFGTITISDAISMGTEGMKASLISREIIADSVELVTFGERMDGIVT  1339

Query  388   IGGCDKNMPGMMMAMIRTDLPSVFLYGGSIMPGEHDGRDVTIVQVFEGVGAYATGDMDAD  567
             IGGCDKNMPGMMMA IRTDLPSVFLYGGSIMPGEHDGR+VTI  VFEGVGA A G+M
Sbjct  1338  IGGCDKNMPGMMMAAIRTDLPSVFLYGGSIMPGEHDGREVTIQNVFEGVGAVADGEMSEG  1159

Query  568   ELDDLERNACPGAGACGGMFTANTMASISEVIGLAPLGSASPPAEEESRYDVARETGELA  747
             ELD++ER+ACPGAG+CGGMFTANTMASISE +G APLGSASPPAE ESRY+ AR  GELA
Sbjct  1158  ELDEMERHACPGAGSCGGMFTANTMASISEALGFAPLGSASPPAEHESRYEEARRAGELA  979

Query  748   VEVIEERRRPSDILTRESFENAIALQTAIGGSTNAVLHLLAMAAEAGVELDIEDFDEISR  927
             VEV++ERR PSD LTRESFENAIALQ A+GGSTNAVLHLLA+AAEAG++LDIE F+EIS
Sbjct  978   VEVVQERRSPSDFLTRESFENAIALQVAVGGSTNAVLHLLALAAEAGIDLDIETFNEISA  799

Query  928   RTPKIADLQPGGESVMNDLHEIGGVPVVLRRLLEADLLHGDAMTITGRTLAEEIEHLE  1101
             RTPKIADLQPGGE VMNDLHE+GGVPVVLR L +A LLHGDA+T+TG T+AEE+E ++
Sbjct  798   RTPKIADLQPGGERVMNDLHEVGGVPVVLRALNDAGLLHGDALTVTGNTIAEELEQID  625
```

Figure 8. The best hit obtained using tBlastx

I then took the missing amino acids from this ortholog from Haloarcula marismortui and blasted this partial gene with the 4 small contigs using tBlastx. The best hit with E-value 6 e -16 was found on contig “Halorhabdus utahensis AX-2, DSM 12940 : HutaDRAFT_4083004_C33”.

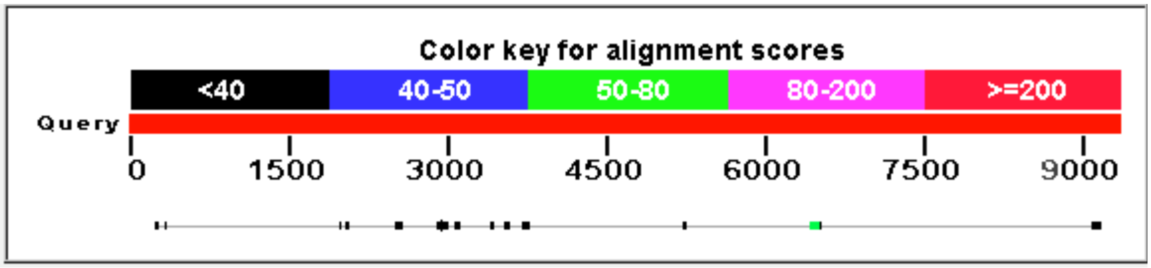


Figure 9. Best hit with the partial gene on “Halorhabdus utahensis AX-2, DSM 12940 : HutaDRAFT_4083004_C33”

However, the aligned position was on the middle of the contig and I could not overlap the sequence of the big contig with this small contig since they were not similar. It was likely that the 3’ end of this gene was lost in the gap of these contigs.

Discussion

The comparison of the three annotations enabled us to understand more about the *H. utahensis* genome. We found that there were significant differences in the results of the three annotation tools suggesting that there is room for improvement. Manatee found more genes because its cut-off value for gene size was smaller. On the other hand, Rast found much fewer genes because it had a higher cut-off value for gene size. Since a lot of the genes Rast missed appeared to be real genes with specific functions, it seems that Rast’s cut-off value was too high.

When we looked at the two Venn diagrams, we noticed that (2764-1471) 1293 ORFs all have the same stop codon across the three annotations, but these annotation tools failed to predict the same start codon. Why did these tools pick up different start condon when it appeared that they all were looking at the same gene? The scale of this discrepancy suggested that gene prediction algorithms were far from perfect and hand curation was still required to complement the tools.

In general, JGI website is the easiest to navigate with its service being relatively more user-friendly. For each gene, JGI provided various links to other relevant databases or tools, making it easy for hand curation. Rast provided metabolic pathway maps and automatically colored all the found EC numbers on the pathway maps. Though, sometimes, a few EC numbers were missing on the maps probably due to software error, Rast is a very good tool especially for studying metabolic pathways of the genomes. Of the three websites, Manatee was the hardest to navigate.

One of the challenges we encountered while comparing the three annotation tools was that for the same gene, all the tools would have different vocabularies for its function. In addition, for the same protein, different tools might call different EC numbers. Each gene prediction algorithm has its own advantage and disadvantage. We will gain more insights for the genome that we study by comparing different tools. Thus, it is important for these tools to adopt gene oncology for defining gene functions making it easier to compare to contrast.

Acknowledgements:

We would like to thank our contacts at JGI, Cheryl Kerfeld and Edwin Kim. We would like to thank Dr. Jonathan Eisen at UC Davis and Dr. Gary Stormo at Washington University for guidance during this project. We would like to thank Matt DeJongh at Hope College for RAST for assisting us in using the RAST server and Ramana Madupu at the J. Craig Venter Institute for doing the same with Manatee database. We would especially like to thank Kjeld Ingvorsen at the Biologisk Insitut at Aarhus Universitet in Denmark for collaborating with us and conducting experiments with *H. utahensis.* Finally, we would like to thank Chris Healey at Davidson College for ordering and growing *H. utahensis* locally.

References:

1. Wainø M, Tindall BJ, and Ingvorsen K (2000) *Halorhabdus utahensis* gen. nov., sp. nov., an aerobic, extremely halophilic member of the *Archaea* from Great Salt Lake, Utah. *International Journal of Systemic and Evolutionary Microbiology* 50:183-190.

2. Hingamp, P. (2008) Metagenome annotation using a distributed grid of undergraduate students. PLOS Biology, 6.11:2362-67