Improving genome start site annotations with multiple alignments of homologous proteins

Derrick E. Wood
dwood@cs.umd.edu
Center for Bioinformatics and Computational Biology
Department of Computer Science
University of Maryland, College Park

Abstract

There is no universal standard for automatic genome annotation, and this has led to a situation where our public repositories of genome annotations contain annotations of differing quality, especially with respect to start site annotation. I present a new method for finding genes, PHANTIM, that uses alignments of protein sequences to homologous proteins, and produces a set of genes for which over 99% of start sites are likely to be correct. Such sets can be used to make corrections to existing annotations and make conclusive statements about the quality of these annotations. Examination of this method's recommended corrections also reveals that some existing annotations, especially those with high usage of rare start codons, are in need of further review. In addition to presenting PHANTIM's results, I make recommendations for annotation pipelines to avoid the types of errors PHANTIM detects.

1 Introduction

The annotation of genomes is a problem for which many computational methods have been devised. However, the "gold standard" for annotation remains experimental verification of results, a time-consuming task that is undertaken for a percentage of genomes that grows smaller as the number of sequenced genomes continues to rise dramatically. Therefore, the use of automated annotation methods will likely continue to be the dominant approach to genome annotation.

A combination of manually reviewed and automatically generated annotations of genomes is available in NCBI's reference sequence database, or RefSeq (Pruitt *et al.*, 2007). Although a portion of the RefSeq database has been manually reviewed by NCBI staff, many genomes' annotations are currently listed as "provisional," meaning that they have not been reviewed and are mostly the same as the annotation submitted by the group that submitted the genome itself. This means that many of the annotations used are from many different sources, are performed by different methods, and are of varying degrees of quality. Yet such annotations are intended to guide genomics research, and many scientists use these annotations without any assurances as to their quality. While some loss of accuracy is almost certain when using an automated method to locate genes, it would be helpful to have an estimate of the confidence one could have in a given annotation.

Gene finding programs typically rely on some combination of coding DNA models, start site signal detection, and comparative methods. To create models of coding DNA, sets of known or likely protein coding regions are used to train the model; the gene finders Glimmer (Salzberg *et al.*, 1998; Delcher *et al.*, 1999), GeneMark.hmm (Lukashin and Borodovsky, 1998), and Prodigal (Hyatt *et al.*, 2010) all work in this fashion. Start sites can be predicted with coding models alone, but their accuracy is rather poor, and so many attempts were made to develop post-processing tools to adjust start site prediction, often focusing on the ribosomal binding site and other signals found upstream of the start codon. Such tools included RBSFinder (Suzek *et al.*, 2001), GS-Finder (Ou, 2004), and TiCO (Tech *et al.*, 2005). Recent programs, including Glimmer version 3 (Delcher *et al.*, 2007), GeneMarkS (Besemer *et al.*, 2001), and Prodigal, have incorporated start site signal recognition into their prediction method to avoid the need for post-processing improvements.

Each of these methods rely little on knowledge of other genomes that have been sequenced, and so can work on new, unknown genomes. As the number of sequenced genomes has increased, however, useful information is available for annotation of both new and existing genomes. Many of the over 1250 bacterial genomes in RefSeq are close relatives of each other. Exploiting the conservation of sequence in these close relatives has been a strategy used to improve genome assembly (Pop *et al.*, 2004) and gene finding (Frishman *et al.*, 1998; Badger and Olsen, 1999). Product hidden Markov models have also been used to improve start site predictions using close genetic relatives (Walker *et al.*, 2002).

Nonetheless, in spite of the large progress made in computational gene prediction, and especially in start site prediction, the accuracy of the methods is difficult to ascertain with certainty. Little experimentally verified data exists for start site locations, and start site prediction accuracy varies from genome to genome. Prodigal appears to have a slight edge at present over the other existing methods of start site prediction, reporting an accuracy of 97% across several genomes, but varying from 91.1% to 98.5%. Many other methods report accuracy in a range from 90% to 96%.

The variable nature of most methods' accuracy is due in

large part to the fact that they must report a start site prediction for every gene in a genome. It is possible, however, to report only a subset of predicted genes for which we can state with high certainty that the start sites are correct. For such genes, comparative evidence can show that the predicted start site is the only one that can explain the sequence conservation seen between genomes. These sets of genes can then be compared to a genome's existing annotation; a high rate of agreement would mean the annotation was highly accurate, while lower agreement could indicate a need to reexamine the annotation.

To this end, I created PHANTIM (Protein Homology-based ANnoTation IMprovement), a tool utilizing multiple alignments of homologous proteins and a set of strict rules that allow only highly accurate predictions to be made. An analysis of PHANTIM's predictions revealed 100% accuracy on a small set of experimentally verified genes, and indicated that its precision on a set of 14 genomes was likely well over 99%. Furthermore, on several genomes, analysis indicated the existing annotations had many start sites that were mislabeled, and were completely missing some genes, many of which could be found with simple homology searches to a database of known proteins.

2 Methods

The goal of PHANTIM is to report a set of genes in a given genome for which there exists evolutionary evidence supporting both the 3′ and 5′ ends of the reported genes. This is done by comparing each predicted gene in a genome with several homologous genes in closely-related genomes, and searching for situations where conservation between the genes implies a necessary protein domain. Where such situations exist, and where they imply an unambiguous start site, PHANTIM will report the gene as a predicted gene. Through this procedure, PHANTIM reports a subset of genes for each genome that has high precision with respect to both 3′ and 5′ ends, and allows for correction of existing annotations.

2.1 Selection of support genomes

PHANTIM begins operation by having a user identify a genome to act upon, called the "target" genome. Then, PHANTIM must select a set of genomes that: (a) have a close evolutionary relationship with the target genome, and (b) are not *too* closely related to any other genome in the set (as well as the target genome). To find such a set of what PHANTIM calls "support" genomes, PHANTIM utilizes the Jaccard distances between genomes calculated as part of the OperonDB project (Pertea *et al.*, 2009). The 1059 genomes used in creating OperonDB are clustered using the furthest-neighbor clustering algorithm supplied by mothur (Schloss *et al.*, 2009), such that no cluster contains two genomes with a Jaccard distance between them that exceeds 0.4. This clustering is performed only once, and can be reused between executions of PHANTIM.

To select the set of support genomes, the following algorithm is used. The set of support genomes, S, begins as an empty set, and all 1059 genomes with known Jaccard distances are placed in a set C that holds all possible support genome candidates. The target genome, along with all genomes in its cluster, are then removed from C. Then the genome g in C with the smallest Jaccard distance to the target genome is added to S; g, along with all of the genomes in its cluster, are removed from C. This step of selection from and removal from C is repeated until either (a) g has a Jaccard distance of more than 0.65 from the target genome, or (b) C is reduced to the empty set.

This selection process is necessary to ensure that the genomes used for comparative purposes in PHANTIM are not too similar to the target genome. If, for example, two genomes of the same species were used, it is quite possible that alignments between genes from these genomes would show identical stretches of intergenic DNA that is identical only because the two genomes have not had the opportunity to mutate and drift apart. My initial attempts at using comparative genomics to improve annotation failed for this very reason: the core assumption of PHANTIM, that conservation will imply functionality, does not hold if the genes being compared are from extremely close relatives. By requiring several more distant relatives to make its decisions, PHANTIM is able to make much more accurate predictions of genes than it would otherwise.

2.2 Locating coding ORFs and maximal length genes

Once the support genomes are identified, possible genes are found in each genome (including the target) by running Glimmer. A modified version of the `g3-iterated.csh` script supplied with Glimmer, designed to avoid gene prediction in certain regions, is run against each chromosome and plasmid found in each genome's GenBank record. PHANTIM reviews the Glimmer predictions for each chromosome and plasmid, and records three items of information for each gene prediction: (a) the coordinates and amino acid translation of the gene's open reading frame (ORF); (b) the coordinates within the ORF of the first two possible start codons codons within the ORF; and (c) the coordinates and amino acid translation of the maximal length gene (the gene using the first possible start codon as a start codon).

Glimmer was used as the gene finder for PHANTIM because of its high sensitivity; as only the genes selected in this step are possible candidates for inclusion in PHANTIM's final report, it is important to have as many true genes found as possible. However, Glimmer's start site predictions are actually ignored by PHANTIM, as all start site predictions for PHANTIM will be made using conservation; only Glimmer's 3′ and respective ORF predictions are carried forward.

There are certain types of genes that Glimmer does not recognize, however. These include genes with a programmed frame shift, selenoproteins, and RNA genes. Glimmer also can not determine if a region that looks like a protein coding gene is actually a pseudogene (a region that used to be a gene but is no longer functional). To avoid a situation where PHANTIM would make predictions that included parts of these kinds of genes (thereby making erroneous predictions), regions that are annotated in GenBank as selenoproteins, pseudogenes, RNA genes, or regions labeled as a gene feature but lack a corresponding CDS feature are excluded

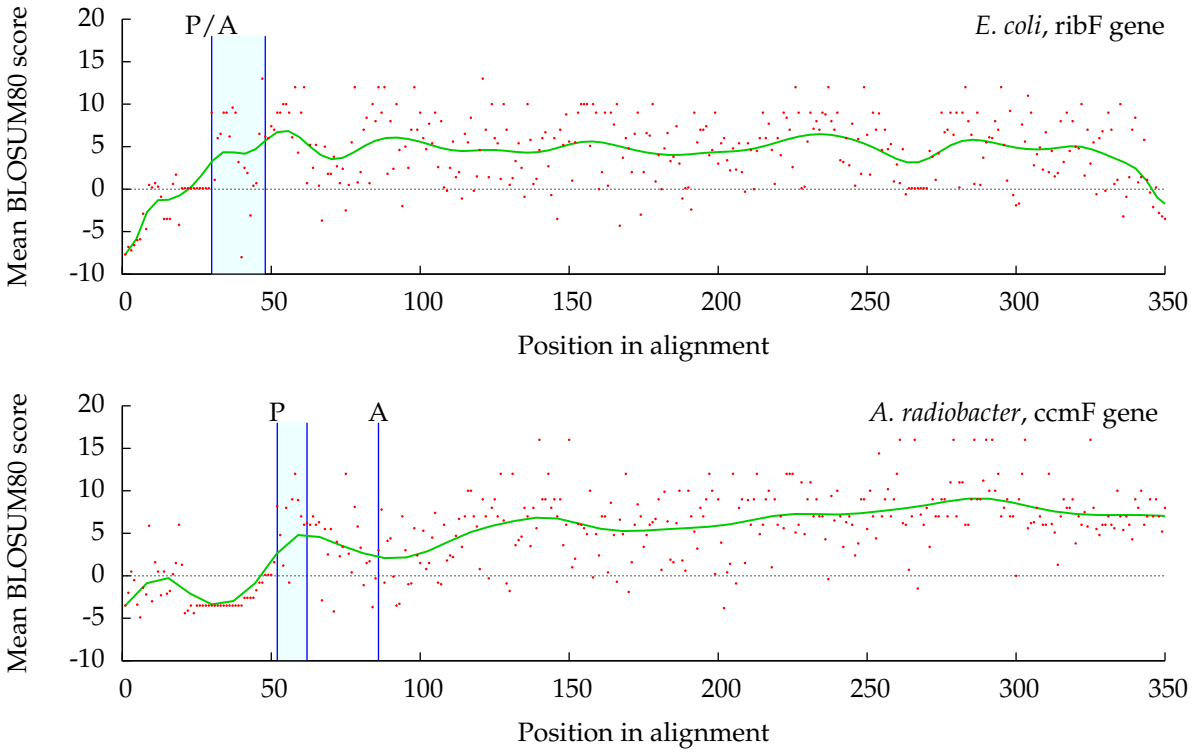

Figure 1. Conservation is shown in alignments of the ribF gene (b0025) in *E. coli* and the ccmF gene (Arad_1495) in *A. radiobacter*, each against 10 homologous genes in different species. The mean BLOSUM80 scores in each column of the alignments are plotted as red dots, with the green line representing the trend of these points. The leftmost blue vertical line in each plot, labeled "P", indicates the first possible start codon, predicted by PHANTIM. The second blue line is the second possible start codon, and the space between it and the first line is the region examined for conservation. The blue line labeled "A" indicates the position of the annotated start in RefSeq; for *E. coli*'s ribF gene, the prediction and annotation agree, and for *A. radiobacter*'s ccmF gene, the annotation is downstream from the prediction. To simplify the figure, "position in alignment" is relative to the beginning of the target gene in each alignment, and only 350 positions are shown of each alignment.

from gene prediction in Glimmer. 50 bp is removed from the edge of each of these excluded regions so that genes that may overlap these regions can still be predicted by Glimmer.

2.3 Finding sets of homologous genes

PHANTIM requires that several homologous genes in support genomes exhibit conservation with a target gene in order to make a prediction about the target gene's coordinates. To find these homologous genes, PHANTIM begins by placing all the maximal length genes from the support genomes into a protein database. BLASTP (Altschul *et al.*, 1990) is then run against this database, using the target genome's maximal length genes as query sequences. As the target and support genomes are intended to be closely related, BLASTP is directed to use the BLOSUM80 substitution matrix (Henikoff and Henikoff, 1992) when scoring its alignments.

The results from BLASTP are filtered in several ways, to increase the likelihood that the support genes used in comparison will be suitable for verifying start site locations. Alignments are discarded if any of the following apply: (a) the alignment's E-value is greater than 10^{-3}; (b) the identity within the alignment is less than 45%; (c) the alignment's length is less than 90% of the target gene's length; or (d) the difference between the length of the support gene in the alignment and the length of the target gene is more than 10% of the length of the target gene. The list of BLASTP hits is further narrowed by ensuring only one gene from any given support genome can be matched with a given target gene.

The remaining results are then used to create sets of homologous genes, with one set per target gene. Each set consists of the top 10 support genes that have alignments with a given target gene in the filtered BLASTP results. To ensure that predictions are made with sufficient evidence, sets with less than 3 support genes are discarded.

2.4 Scoring multiple alignments

For each set of homologous support genes, the genes' respective ORF translations are placed into a multiple alignment along with the ORF translation of the set's corresponding target gene, using MUSCLE (Edgar, 2004). PHANTIM then examines the window of columns in the multiple alignment that represent the first (inclusive) and second (exclusive) possible start codons in the target gene. This region is highlighted in Figure 1. Within this window, each column is examined separately, and the amino acids in the support genes are scored by their similarity to the target gene's amino acid within that

column, using the BLOSUM80 matrix (gap characters receive a score of -8 for non-identity substitution, and +1 for alignment of two gap characters). For each column, the mean of these scores is calculated, and then the mean of the column scores throughout the window is found. If this mean substitution score is at least 3.0, and the length of the window is at least 7 amino acids, PHANTIM will declare the first possible start codon to be the correct one and report the gene as a predicted gene; if these two conditions do not hold, no prediction whatsoever is made regarding the gene.

Due to the fact that the similarity between homologous genes can break down toward the 5′ ends (Delcher *et al.*, 2007), a lack of conservation in a given segment of the target gene does not imply that that segment is not actually part of the gene. It is not possible, though, to make a claim one way or another regarding the gene's start site, and so PHANTIM does not do so in the absence of conservation. Similarly, only the gap between the first and second possible start codons is searched; conservation here does indeed lend support to the claim that the gap is part of a functional protein. But a lack of conservation in that gap, and the presence of conservation in a later part of the gene does not allow PHANTIM to make any firm conclusions as to the correct start site, due to the possibility of the target gene possessing a novel mutation not present in the support genes.

2.5 Running PHANTIM

PHANTIM is designed for use with a Linux operating system, and makes extensive use of Perl and Make, along with other standard Unix utilities, in addition to the various programs cited above. To run PHANTIM with a specific target genome, the name of that genome must be specified. All chromosomes and plasmids in the target genome's GenBank record will have their sequences analyzed for genes that can be annotated with high confidence, and a separate report will be made for each sequence. These reports can then be compared with the GenBank or RefSeq annotation to determine if changes should be made in the current annotation, and all alignments used in making PHANTIM's predictions are retained for possible manual examination.

As the number of support genomes increases, the amount of gene finding and the size of the BLASTP database increase as well, leading to long running times. To reduce overall execution time, PHANTIM is designed to use (by default) all processing units on a computer and store gene finding results for use in future executions.

2.6 Evaluation

To evaluate PHANTIM, it was run against 13 bacterial genomes: *Acidovorax citrulli* AAC00-1, *Agrobacterium radiobacter* K84, *Bacillus anthracis* Ames, *Bacillus subtilis* 168, *Bradyrhizobium* BTAi1, *Chromobacterium violaceum* ATCC 12472, *Escherichia coli* K12 substr. MG1655, *Helicobacter pylori* 26695, *Mycobacterium tuberculosis* CDC1551, *Neisseria meningitidis* MC58, *Staphylococcus aureus* MRSA252, *Vibrio cholerae* El Tor N16961, and *Xenorhabdus bovienii* SS 2004. As the Jaccard distances between genomes were based off of the sequences stored in GenBank, the GenBank sequences for these genomes, and all support genomes, were used during execution. Prediction sets were compared to the corresponding annotations stored in RefSeq; the results of these comparisons are shown in Table 1. Predictions that did not match a RefSeq annotated coding sequence's 5′ and 3′ ends were examined manually, by inspection of the prediction's corresponding alignment. In the case of predictions without a matching 3′ end, BLASTP was run against the non-redundant database (limited to bacteria only), with the predicted gene and any overlapping annotated genes used as query sequences.

Conclusively evaluating the start-site prediction accuracy of PHANTIM is difficult due to the lack of experimentally-verified data about start sites. For *E. coli*, there exists a large set of genes for which the start sites have been verified by N-terminal sequencing. 878 of these genes have been documented in the EcoGene database (Rudd, 2000), and comparison to these genes were also used in addition to the RefSeq annotations to measure PHANTIM's accuracy.

3 Results and Discussion

3.1 High agreement in both 3′ and 5′ predictions

Among the genomes examined, only with one, *M. tuberculosis*, was the percentage of predictions with a 3′ end match in the RefSeq annotation less than 99%. All genomes also had at least 92% agreement between predictions and 5′ end annotations. Genomes with higher GC content tended to have lower 5′ agreement, likely a consequence of the difficulty in discerning the correct start codon in longer ORFs. In total, of the 6801 predicted genes, 97.5% had 5′ and 3′ matches.

Given the high agreement between predictions and the RefSeq annotations, it is likely that PHANTIM yields a set of high-precision predictions. Further evidence of PHANTIM's high 5′ prediction precision is shown when the predictions of *E. coli* genes are compared to the known start sites in the experimentally verified set of 878 genes found in the EcoGene database. Of these 878 genes, 288 (32.8%) were predicted by PHANTIM, and all 288 matched a gene in the verified set on both the 5′ and 3′ ends. That the 3′ sensitivity of PHANTIM against this verified set is much higher than against the full RefSeq annotation may be indicative of a large number of incorrectly annotated hypothetical proteins, or of a bias in the EcoGene set toward proteins that use the 5′-most start codon.

Finally, upon manual examination (see Appendices A and B), all of the 3′ disagreements resulted in the addition of new genes. 131 of the 155 disagreements were resolved in favor of PHANTIM, and only 3 were clearly resolved in favor of the RefSeq annotation; each of these 3 were due to the use of a rare ATT or CTG start codon for the gene. PHANTIM's high percentage of of predictions that either agree with the existing annotation, or are verified by examination, gives strong support to the idea that it provides a high-precision set of start-site predictions, and it also provides a rough estimate as to how precise the predictions are. With over 99.5% of predictions validated by agreement or examination, the predictions of PHANTIM should be regarded as highly correct, and a large number of annotated genes that conflict with its predictions should be indicative of a lower-quality annotation.

Table 1. Comparison of PHANTIM predictions to RefSeq annotations

Genome			Gene Counts			Matches with RefSeq Annotation			
Organism	GC%	SG	RefSeq	Pred.	%	3′ Matches		5′ & 3′ Matches	
A. citrulli	69	56	4709	466	10	465	99.8%	457	98.1%
A. radiobacter	60	63	6107	570	9	565	99.1%	529	92.8%
B. anthracis	35	19	5328	698	13	697	99.9%	695	99.6%
B. subtilis	44	40	4176	578	14	578	100.0%	574	99.3%
Bradyrhizobium BTAi1	65	48	7393	679	9	678	99.9%	633	93.4%
C. violaceum	65	70	4407	435	10	435	100.0%	420	96.6%
E. coli	51	65	4145	789	19	789	100.0%	783	99.2%
H. pylori	39	9	1573	197	13	197	100.0%	195	99.0%
M. tuberculosis	66	16	4189	260	6	257	98.9%	249	95.8%
N. meningitidis	52	28	2063	219	11	219	100.0%	219	100.0%
S. aureus	33	31	2650	352	13	352	100.0%	352	100.0%
V. cholerae	47	52	3834	711	19	710	99.9%	710	99.9%
X. bovienii	45	24	4260	847	20	847	100.0%	818	96.6%
		Totals	54834	6801	12	6789	99.8%	6634	97.5%

"SG" is the number of support genomes used. 3′ matches are predicted genes that share a stop codon with a CDS in the RefSeq annotation. 5′ & 3′ matches are predicted genes that share both a start and stop codon with an annotated CDS.

Table 2. Results of examination of 5′ differences in prediction

Genome		PHANTIM comparison		Recommendations				
Organism	GC%	3′ matches	5′ mismatches	Change	Review	Keep	5′ matched/verified	
A. citrulli	69	465	8	6	2	0	463	99.6%
A. radiobacter	60	565	36	33	3	0	562	99.5%
B. anthracis	35	697	2	2	0	0	697	100.0%
B. subtilis	44	578	4	3	0	1	577	99.8%
Bradyrhizobium BTAi1	65	678	45	42	3	0	675	99.6%
C. violaceum	65	435	15	12	3	0	432	99.3%
E. coli	51	789	6	3	1	2	786	99.6%
H. pylori	39	197	2	2	0	0	197	100.0%
M. tuberculosis	66	257	8	6	2	0	255	99.2%
N. meningitidis	52	219	0	0	0	0	219	100.0%
S. aureus	33	352	0	0	0	0	352	100.0%
V. cholerae	47	710	0	0	0	0	710	100.0%
X. bovienii	45	847	29	22	7	0	840	99.2%
	Totals	6789	155	131	21	3	6765	99.6%

All PHANTIM predictions were compared to the respective RefSeq annotation, and for those genes that had a 3′ match but a 5′ mismatch, the alignment was analyzed. Recommendations as a result of that analysis are given here, with "Change" meaning an annotation should be changed, "Review" indicating further review is needed, and "Keep" meaning that the annotation should be left as is. "5′ matched/verified" is a count of predicted genes with 3′ matches and a 5′ end that either matched a RefSeq gene or was verified by examination, along with the percentage of 3′ matches such a count constitutes.

3.2 Discovery of previously unannotated genes

Twelve genes predicted by PHANTIM lack a 3′ match with a gene in RefSeq's annotation. An examination of each of these revealed a very strong likelihood that all twelve of these genes should be added to their respective annotations, and in many cases, replace annotated genes that are strongly overlapped by these predicted genes. Appendix A contains a summary of the recommended changes and the evidence that exists for making them.

Six of these genes exist in regions that are currently annotated as intergenic in RefSeq, or are overlapped by less than 5 nt by another gene. A simpler method of extracting these regions, supplying them as input to TBLASTN, and searching against a database of known bacterial proteins would have found as much evidence as PHANTIM did for these genes, if not far more. Such a method would also find such evidence much faster, and almost certainly would find even more unannotated genes than PHANTIM. Given the large number of bacterial genomes that have been sequenced and annotated, such a process should now be used by any annotation pipeline.

Table 3. Relationship between GC content, high annotation of rare start codons, and PHANTIM disagreement

Genome		RefSeq start codon usage (%)				PHANTIM predictions	
Organism	GC%	CTG	ATT	ATC	ATA	5′ mismatches	Changes
A. citrulli	69	0.00	0.00	0.00	0.00	8	6
A. radiobacter	60	3.95	0.00	0.00	0.00	36	33
B. anthracis	35	0.00	0.00	0.00	0.00	2	2
B. subtilis	44	0.14	0.22	0.05	0.00	4	4
Bradyrhizobium BTAi1	65	0.72	0.00	0.00	0.00	45	42
C. violaceum	65	4.29	0.00	0.00	0.00	15	12
E. coli	51	0.05	0.05	0.00	0.00	6	3
H. pylori	39	0.06	0.06	0.00	0.00	2	2
M. tuberculosis	66	0.00	0.00	0.00	0.00	8	6
N. meningitidis	52	0.00	0.00	0.00	0.00	0	0
S. aureus	33	0.00	0.00	0.00	0.00	0	0
V. cholerae	47	0.00	0.00	0.00	0.00	0	0
X. bovienii	45	3.92	0.00	0.00	0.00	29	22

RefSeq start codon usage is the percentage of annotated genes using the given start codon. 5′ mismatches are instances where the prediction had a stop codon match with the annotation, but not the start codon. Changes are the number of 5′ mismatches resolved in favor of the prediction after manual examination.

Six other predicted genes have large overlaps with annotated genes, in most cases being completely overlapped by the annotated gene. In each of these six cases, the predicted gene has far more alignments against bacterial proteins in NCBI's non-redundant protein database, nr, than does the corresponding overlapping annotated gene. This larger amount of sequence conservation in other bacteria found in the predicted genes serves as strong evidence of the correctness of these predictions; in light of the high overlap in these situations, the higher conservation also serves as evidence of the incorrectness of the conflicting annotations.

Like those genes found in regions believed to be intergenic, the predicted genes that were overlapped were all found as part of the Glimmer prediction set. Even in the face of a conflicting annotation from another source, the conflict can be resolved by the use of a BLAST search against other genomes. In this case, a TBLASTN search of two overlapping potential genes against all bacterial genomes would provide guidance as to which of the two genes was more conserved, and thus more likely to be a true gene. Once again, with the rise in sequenced bacterial genomes in GenBank, there exist considerable resources for such a method, and it would be a useful addition to an annotation pipeline.

3.3 Extensions to 5′ ends of genes

The primary motivation behind PHANTIM was to correct start site annotations, and it performs quite well at this task. Review of the alignments in those cases where PHANTIM's 5′ prediction of a gene disagreed with RefSeq's annotation revealed that in nearly all cases, the PHANTIM prediction had substantial evidence in favor of it; for each genome, over 99% of start site predictions either matched the RefSeq annotation or were verified by examination. The full results of this evaluation are in Appendix B, and are summarized in Table 2. This high percentage of gene predictions that agree with the annotation or have been verified supports the idea that PHANTIM provides a high-precision set of start site predictions, and it also provides a rough estimate as to exactly how precise the predictions are.

With the exception of genes annotated with a rare start codon, the gene predicted by PHANTIM is always at least as long as the corresponding annotated gene, as should be expected given that PHANTIM will only predict a start site that lies at the 5′-most possible start codon. In some cases, PHANTIM will find regions of more than 100 nt that have been wrongly omitted from the 5′ end of a gene's annotation. With the exception of *X. bovienii*, in most of the low (less than 60%) GC content genomes, the number of genes that had changes made were very low, and were less than 1% of the genes predicted by PHANTIM. As might be expected, this was not the case for high-GC genomes, which are more prone to start site misannotation.

3.4 High erroneous annotation of rare start codons

There appears to be a connection between the quality of an annotation and the number of rare start codons used in the annotation. By default, Glimmer will only predict ATG, GTG, and TTG start codons; the NCBI genetic code for bacteria also permits translation initiation with CTG, ATT, ATC, and ATA codons. These four codons are rarely used, however, and are often not considered by gene finders. In two highly-studied genomes, *B. subtilis* and *E. coli*, a total of 8321 genes are annotated in RefSeq; only 21 (0.25%) use one of these four rare start codons. Many other genomes have zero usage of rare start codons in their RefSeq annotations.

While using a significantly higher proportion of rare start codons than 0.25% would not be impossible for another genome, such usage would require significant evidence to justify it. Of the 13 genomes used in PHANTIM's evaluation, four had rare start codon levels in excess of 0.7%, and of the rare start codons, these four exclusively used CTG. Each of these four genomes had high numbers of 5′ mismatches as well as recommended changes when compared to the other genomes examined (Table 3). Although three of these genomes are

Table 4. Results of running PHANTIM on five genomes with high rare start codon usage

Genome			RefSeq start codon usage (%)				Gene counts		Matches with RefSeq annotation			
Organism	GC%	SG	CTG	ATT	ATC	ATA	RefSeq	Pred.	3′ Matches		5′ & 3′ Matches	
C. turicensis z3032	57	54	8.36	5.06	3.06	4.51	4213	938	936	99.8%	773	82.4%
M. hyopneumoniae 7448	28	7	1.37	17.81	7.91	3.65	657	56	28	50.0%	20	35.7%
M. hyopneumoniae J	29	7	1.37	17.96	8.68	3.96	657	54	27	50.0%	18	33.3%
M. synoviae 53	29	10	0.46	13.51	9.71	2.28	659	95	36	37.9%	21	22.1%
R. massiliae MTU5	33	7	2.58	19.94	23.14	4.65	968	78	78	100.0%	37	47.4%

Column headings are the same as in Tables 1 and 3. Note that the three *Mycoplasma* genomes (*M. hyopneumoniae* strains 7448 and J, and *M. synoviae*) use the usual stop codon TGA to encode for tryptophan, a deviation from the standard genetic code that is partly responsible for the low 3′ agreement between PHANTIM and RefSeq.

also high in GC content, the *X. bovienii* genome is only 45% GC, and so it does not appear high GC content is responsible for the high disagreement between PHANTIM and these RefSeq annotations. The only common factor I have found between these genomes and their annotations is a higher-than-expected proportion of CTG start codons. Curiously, there does not appear to be any documented reason for such a high usage of this rare start codon for any of these four genomes (Brazilian National Genome Project Consortium, 2003; Slater *et al.*, 2009; Giraud *et al.*, 2007).

The high rate of PHANTIM disagreement over a small subset of genes is likely be indicative of a weakness in 5′ annotation in the methods used to annotate these genomes. In particular, the ability of these methods to automatically call a CTG codon the start of a gene has in many cases caused the gene to be annotated incorrectly. These results suggest that any pipeline seeking to call CTG start codons should only do so when there is considerable evidence in favor of it. Furthermore, the high annotation of CTG start codons, or other rare start codons, may be a simple indicator that an existing annotation is in need of review.

An inspection of the bacterial chromosome annotations in RefSeq revealed five genomes that used rare start codons for more than 5% of genes, and had at least 50 such genes. I ran PHANTIM on each, and found extraordinarily high disagreement between the annotations and PHANTIM's predictions, not only with regard to 5′ predictions, but in some cases, 3′ predictions as well (see Table 4). An inspection of the 3′ disagreements revealed many were due to the use of the normal stop codon TGA to encode tryptophan, a deviation from the standard genetic code known to occur in some *Mycoplasma* species (Inamine *et al.*, 1990). This property of *Mycoplasma* makes genomes difficult to annotate properly, and it also makes analysis of PHANTIM's results difficult as well. However, of the eight 5′ disagreements for *M. hyopneumoniae* 7448, 3 of the annotated genes were annotated with an ATT start codon, in spite of the presence of an in-frame ATG start codon within 18 nt of the ATT codon. Such a preference for a rare start codon over a nearby standard start is something that would be quite remarkable, and yet there is no documented justification of this preference, nor even a mention of it for *M. hyopneumoniae* and *M. synoviae* (Vasconcelos *et al.*, 2005). Inspection of the 5′ disagreements of *C. turicensis* and *R. massiliae* both showed a large number of corrections that would be made to genes annotated with rare start codons; once again, there is no documented justification for these genomes' annotations' high use of such rare start codons (Stephan *et al.*, 2011; Blanc *et al.*, 2007).

3.5 Factors affecting number of predictions

High GC genomes have ORFs that extend farther upstream from the true start codon than do low GC genomes; this region upstream of the true start also contains more possible start codons in high GC genomes. As can be seen in Table 1, high GC genomes have a lower number of predictions made as a percentage of total genes than do lower GC genomes. This is due to two factors associated with a longer upstream region. First, the extra possible start codons can cause the maximal length genes to be of widely differing sizes in different genomes; PHANTIM's requirement for homologous maximal length genes to be of approximately the same length can eliminate possible useful homologs from consideration. This in turn can lead to a target gene not having enough homologs with which to have an alignment, which will lower the number of predicted genes. Another factor that drives prediction count down in high GC genomes is that PHANTIM will only predict genes where it finds evidence that the first possible start codon is the correct one; obviously, the more possible start codons upstream of true start codons that a genome has, the fewer genes that can be predicted by such a method.

In addition, low numbers of support genomes can result in a very low number of predictions. This occurs when an organism has not had many of its closer evolutionary neighbors sequenced. To aid in further comparative study of such organisms, attention should be paid to sequencing some of these areas of the bacterial tree of life.

4 Conclusion

PHANTIM's gene predictions are quite accurate, with a precision that likely exceeds 99%. Such high precision enables it to make corrections to existing annotations. Application of PHANTIM to some existing annotations reveals that high usage of rare start codons implies a high error rate in start site annotation. A small number of genes that were omitted from annotations were also discovered by PHANTIM; half of these would have been discovered by a BLAST search of the annotated intergenic regions.

The results of comparing PHANTIM's predictions with RefSeq's annotations reveal that researchers seeking to use these annotations should be aware that although these annotations are mostly correct, they are not completely accurate. Researchers should also be cautious of start site annotation accuracy, especially for annotations with high rare start codon usage. Those who annotate genomes should also ensure that the pipelines they use to perform their work, and the manner in which they use these pipelines, present results that are not highly inconsistent with known genomes; if they are so inconsistent, further — or at least some — justification should be given for this high deviation from the norm.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Badger, J. H. and Olsen, G. J. (1999). CRITICA: coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution*, **16**(4), 512–524.

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, **29**(12), 2607–2618.

Blanc, G., Ogata, H., Robert, C., Audic, S., Claverie, J.-M. M., and Raoult, D. (2007). Lateral gene transfer between obligate intracellular bacteria: evidence from the Rickettsia massiliae genome. *Genome Research*, **17**(11), 1657–1664.

Brazilian National Genome Project Consortium (2003). The complete genome sequence of Chromobacterium violaceum reveals remarkable and exploitable bacterial adaptability. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(20), 11660–11665.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, **27**(23), 4636–4641.

Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**(6), 673–679.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797.

Frishman, D., Mironov, A., Mewes, H.-W., and Gelfand, M. (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Research*, **26**(12), 2941–2947.

Giraud, E., Moulin, L., Vallenet, D., Barbe, V., Cytryn, E., Avarre, J.-C. C., Jaubert, M., Simon, D., Cartieaux, F., Prin, Y., Bena, G., Hannibal, L., Fardoux, J., Kojadinovic, M., Vuillet, L., Lajus, A., Cruveiller, S., Rouy, Z., Mangenot, S., Segurens, B., Dossat, C., Franck, W. L., Chang, W.-S. S., Saunders, E., Bruce, D., Richardson, P., Normand, P., Dreyfus, B., Pignol, D., Stacey, G., Emerich, D., Verméglio, A., Médigue, C., and Sadowsky, M. (2007). Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia. *Science*, **316**(5829), 1307–1312.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**(22), 10915–10919.

Hyatt, D., Chen, G. L., LoCascio, P., Land, M., Larimer, F., and Hauser, L. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**(1), 119+.

Inamine, J. M., Ho, K. C., Loechel, S., and Hu, P. C. (1990). Evidence that UGA is read as a tryptophan codon rather than as a stop codon by Mycoplasma pneumoniae, Mycoplasma genitalium, and Mycoplasma gallisepticum. *Journal of Bacteriology*, **172**(1), 504–506.

Lukashin, A. V. and Borodovsky, M. (1998). GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research*, **26**(4), 1107–1115.

Ou, H. (2004). GS-Finder: a program to find bacterial gene start sites with a self-training method. *The International Journal of Biochemistry & Cell Biology*, **36**(3), 535–544.

Pertea, M., Ayanbule, K., Smedinghoff, M., and Salzberg, S. L. (2009). OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Research*, **37**(suppl 1), D479–D482.

Pop, M., Phillippy, A., Delcher, A. L., and Salzberg, S. L. (2004). Comparative genome assembly. *Briefings in Bioinformatics*, **5**(3), 237–248.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **35**(Database issue), D61–D65.

Rudd, K. E. (2000). EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Research*, **28**(1), 60–64.

Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, **26**(2), 544–548.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**(23), 7537–7541.

Slater, S. C., Goldman, B. S., Goodner, B., Setubal, J. a. C., Farrand, S. K., Nester, E. W., Burr, T. J., Banta, L., Dickerman, A. W., Paulsen, I., Otten, L., Suen, G., Welch, R., Almeida, N. F., Arnold, F., Burton, O. T., Du, Z., Ewing, A., Godsy, E., Heisel, S., Houmiel, K. L., Jhaveri, J., Lu, J., Miller, N. M., Norton, S., Chen, Q., Phoolcharoen, W., Ohlin, V., Ondrusek, D., Pride, N., Stricklin, S. L., Sun, J., Wheeler, C., Wilson, L., Zhu, H., and Wood, D. W. (2009). Genome sequences of three agrobacterium biovars help elucidate the evolution of multichromosome genomes in bacteria. *Journal of Bacteriology*, **191**(8), 2501–2511.

Stephan, R., Lehner, A., Tischler, P., and Rattei, T. (2011). Complete genome sequence of Cronobacter turicensis LMG 23827, a food-borne pathogen causing deaths in neonates. *Journal of Bacteriology*, **193**(1), 309–310.

Suzek, B. E., Ermolaeva, M. D., Schreiber, M., and Salzberg, S. L. (2001). A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**(12), 1123–1130.

Tech, M., Pfeifer, N., Morgenstern, B., and Meinicke, P. (2005). TICO: a tool for improving predictions of prokaryotic translation initiation sites. *Bioinformatics*, **21**(17), 3568–3569.

Vasconcelos, A. T. R., Ferreira, H. B., Bizarro, C. V., Bonatto, S. L., Carvalho, M. O., Pinto, P. M., Almeida, D. F., Almeida, L. G. P., Almeida, R., Alves-Filho, L., Assuncao, E. N., Azevedo, V. A. C., Bogo, M. R., Brigido, M. M., Brocchi, M., Burity, H. A., Camargo, A. A., Camargo, S. S., Carepo, M. S., Carraro, D. M., de Mattos Cascardo, J. C., Castro, L. A., Cavalcanti, G., Chemale, G., Collevatti, R. G., Cunha, C. W., Dallagiovanna, B., Dambros, B. P., Dellagostin, O. A., Falcao, C., Fantinatti-Garboggini, F., Felipe, M. S. S., Fiorentin, L., Franco, G. R., Freitas, N. S. A., Frias, D., Grangeiro, T. B., Grisard, E. C., Guimaraes, C. T., Hungria, M., Jardim, S. N., Krieger, M. A., Laurino, J. P., Lima, L. F. A., Lopes, M. I., Loreto, E. L. S., Madeira, H. M. F., Manfio, G. P., Maranhao, A. Q., Martinkovics, C. T., Medeiros, S. R. B., Moreira, M. A. M., Neiva, M., Ramalho-Neto, C. E., Nicolas, M. F., Oliveira, S. C., Paixao, R. F. C., Pedrosa, F. O., Pena, S. D. J., Pereira, M., Pereira-Ferrari, L., Piffer, I., Pinto, L. S., Potrich, D. P., Salim, A. C. M., Santos, F. R., Schmitt, R., Schneider, M. P. C., Schrank, A., Schrank, I. S., Schuck, A. F., Seuanez, H. N., Silva, D. W., Silva, R., Silva, S. C., Soares, C. M. A., Souza, K. R. L., Souza, R. C., Staats, C. C., Steffens, M. B. R., Teixeira, S. M. R., Urmenyi, T. P., Vainstein, M. H., Zuccherato, L. W., Simpson, A. J. G., and Zaha, A. (2005). Swine and Poultry Pathogens: the Complete Genome Sequences of Two Strains of Mycoplasma hyopneumoniae and a Strain of Mycoplasma synoviae. *Journal of Bacteriology*, **187**(16), 5568–5577.

Walker, M., Pavlovic, V., and Kasif, S. (2002). A comparative genomic method for computational identification of prokaryotic translation initiation sites. *Nucleic Acids Research*, **30**(14), 3181–3191.

A Examination of PHANTIM predictions lacking 3′ matches in RefSeq

This appendix contains a list of all PHANTIM predictions that do not have a 3′ match in a genome's RefSeq annotation, as well as a recommendation regarding any changes in the annotation with respect to the predicted gene.

Genome	Start	Stop	Str.	Recommendation
A. citrulli	2978649	2978326	rev.	**Add** gene to annotation. This gene is an exact nucleotide copy of Aave_2936. This gene only overlaps one gene, Aave_2709, by 4 nt.
A. radiobacter chr. 1	579295	579161	rev.	**Add** gene to annotation. This gene has over 100 BLASTP hits in nr, all to ribosomal proteins; one hit is a 94% amino acid identity alignment, with 100% coverage, to a "50S ribosomal protein" in *Methylobacterium populi* BJ001. Its region is intergenic according to the RefSeq annotation.
A. radiobacter chr. 1	1323593	1323829	fwd.	**Replace** Arad_1670. This gene is almost completely overlapped by Arad_1670, named an "acyl carrier protein". Arad_1670 has only 26 BLASTP hits in nr, and only 9 with E-values less than 10^{-3}. In comparison, this region has over 100 BLASTP hits, with the 100th best having an E-value of 5e-21; the name of all these hits is also "acyl carrier protein".
A. radiobacter chr. 1	1571695	1572003	fwd.	**Add** gene to annotation. This gene has over 100 BLASTP hits to ribosomal proteins in nr, with 100% amino acid identity to a "30S ribosomal protein S10" in *Rhizobium etli*. Its region is intergenic according to the RefSeq annotation.
A. radiobacter chr. 1	1669050	1668718	rev.	**Replace** Arad_2117. 77 nt overlap with Arad_2117, a hypothetical protein with only 3 BLASTP hits in nr, the best having an E-value of 2.1. This gene has over 100 BLASTP hits in nr, one with 88% amino acid identity to an "iron-sulfur cluster assembly accessory protein" in *Rhizobium leguminosarum*.
A. radiobacter chr. 2	1069411	1069803	fwd.	**Add** gene to annotation. This gene has 101 BLASTP hits to genes in nr, including an 87% amino acid identity alignment to a "glutathione-dependent formaldehyde-activating, GFA" gene in *Mesothizobium* sp. BNCI. Its region is intergenic according to the RefSeq annotation.
B. anthracis	3825729	3825601	rev.	**Replace** BA_4175. This gene is overlapped completely by BA_4175, which is a hypothetical protein with only 5 BLASTP hits in nr. This gene has 22 BLASTP hits to genes in various *Bacillus* and other genomes, including a 72% amino acid identity alignment to a phosphoesterase gene in *B. thuringiensis*. In addition, this gene has 81% amino acid identity to BA_4174, located just upstream.
Bradyrhizobium BTAi1	1946535	1946260	rev.	**Add** gene to annotation. This gene has 97 BLASTP hits in nr, most of which are alignments to conserved hypothetical proteins and proteins of unknown function (DUF1153). Its region is currently intergenic according to the RefSeq annotation.
M. tuberculosis	2765730	2765107	rev.	**Replace** MT2541. This gene has over 100 BLASTP hits in nr, including one with 57% amino acid identity to "DSBA oxidoreductase" in *Streptomyces* sp. SPB78. MT2541 is a hypothetical protein that completely overlaps this gene, and has only 17 BLASTP hits, only one of which has an E-value below 1.
M. tuberculosis	2943593	2943240	rev.	**Replace** MT2694. This gene has over 100 BLASTP hits in nr, including one with 98% coverage and 58% amino acid identity to a "Cupin 2 conserved barrel domain-containing protein" in *Nakamurella multipartita*. MT2694 is a hypothetical protein that completely overlaps this gene, and has only 2 BLASTP hits, with E-values of 2.3 and 7.3.
M. tuberculosis	4103085	4103603	fwd.	**Replace** MT3770. This gene has over 100 BLASTP hits in nr, including one with 100% amino acid identity to a "transmembrane protein" in *M. tuberculosis* H37Rv. MT3770 is a hypothetical protein that overlaps all but 72 nt of this gene, and has only 6 BLASTP hits, only one of which has an E-value below 1.

Continued on next page

Genome	Start	Stop	Str.	Recommendation
V. cholerae chr. I	1651882	1652085	fwd.	**Add** gene to annotation. This gene has over 100 BLASTP hits to genes in nr, two of which are "Zn-ribbon proteins" in *Idiomarina* genomes. Its region is intergenic according to the RefSeq annotation.

B Examination of PHANTIM predictions lacking 5′ & 3′ matches in RefSeq

This appendix contains a list of all PHANTIM predictions that have a 3′ match but not a 5′ match in a genome's RefSeq annotation, as well as notes regarding the conservation found in the predicted gene's multiple alignment. Evaluations indicate one or more of the following holds for the gene:

(a) **Change**: strong, clear conservation exists and supports the prediction; annotation should be changed

(b) **ReviewWeak**: gene needs further examination, due to weaker evidence for the change

(c) **ReviewUp**: gene needs further examination, as conservation exists upstream of predicted start (not just the annotated start)

(d) **CTG** or **ATT**: a rare start codon (CTG or ATT) is annotated for this gene

(e) **Keep**: maintain current annotation

		RefSeq			PHANTIM		
Genome	Gene ID	Start	Stop	Str.	Start	Extra (nt)	Evaluation
A. citrulli	Aave_0289	320917	321636	fwd.	320743	+174	**Change**
A. citrulli	Aave_0691	750181	752169	fwd.	749971	+210	**ReviewUp**
A. citrulli	Aave_2611	2860437	2863607	fwd.	2860347	+90	**Change**
A. citrulli	Aave_2652	2914135	2914890	fwd.	2914012	+123	**Change**
A. citrulli	Aave_3114	3439123	3439587	fwd.	3439066	+57	**Change**
A. citrulli	Aave_3366	3722691	3721507	rev.	3722721	+30	**ReviewUp**
A. citrulli	Aave_4383	4876301	4875558	rev.	4876328	+27	**Change**
A. citrulli	Aave_4645	5170949	5170674	rev.	5171006	+57	**Change**
A. radiobacter chr. 1	Arad_0252	225153	224752	rev.	225234	+81	**Change**
A. radiobacter chr. 1	Arad_0288	257347	256451	rev.	257692	+345	**ReviewWeak**
A. radiobacter chr. 1	Arad_0306	270939	272090	fwd.	270489	+450	**ReviewUp**
A. radiobacter chr. 1	Arad_0597	493707	494546	fwd.	493656	+51	**Change**
A. radiobacter chr. 1	Arad_0772	635281	635916	fwd.	635233	+48	**Change CTG**
A. radiobacter chr. 1	Arad_1495	1187396	1189285	fwd.	1187294	+102	**Change**
A. radiobacter chr. 1	Arad_1676	1329161	1328331	rev.	1329158	-3	**Change CTG**
A. radiobacter chr. 1	Arad_1810	1431875	1431402	rev.	1431869	-6	**Change CTG**
A. radiobacter chr. 1	Arad_1862	1475491	1477029	fwd.	1475518	-27	**Change CTG**
A. radiobacter chr. 1	Arad_1909	1517403	1518875	fwd.	1517394	+9	**Change CTG**
A. radiobacter chr. 1	Arad_1938	1541374	1542411	fwd.	1541383	-9	**ReviewUp CTG**
A. radiobacter chr. 1	Arad_2459	1945856	1945128	rev.	1945898	+42	**Change**
A. radiobacter chr. 1	Arad_2513	1986301	1987098	fwd.	1986265	+36	**Change**
A. radiobacter chr. 1	Arad_2726	2165484	2165020	rev.	2165508	+24	**Change CTG**
A. radiobacter chr. 1	Arad_2787	2210814	2209909	rev.	2210859	+45	**Change**
A. radiobacter chr. 1	Arad_3282	2619023	2619982	fwd.	2619059	-36	**Change CTG**
A. radiobacter chr. 1	Arad_3401	2727288	2726584	rev.	2727369	+81	**Change**
A. radiobacter chr. 1	Arad_3755	2996226	2995261	rev.	2996292	+66	**Change**
A. radiobacter chr. 1	Arad_4306	3442004	3440547	rev.	3441977	-27	**Change CTG**
A. radiobacter chr. 1	Arad_4308	3444447	3442915	rev.	3444444	-3	**Change CTG**
A. radiobacter chr. 1	Arad_4366	3491052	3490612	rev.	3491148	+96	**Change**
A. radiobacter chr. 1	Arad_4394	3516274	3513554	rev.	3516244	-30	**Change CTG**
A. radiobacter chr. 1	Arad_4606	3694076	3694975	fwd.	3694103	-27	**Change CTG**
A. radiobacter chr. 1	Arad_4853	3926590	3926063	rev.	3926629	+39	**Change**

Continued on next page

		RefSeq			PHANTIM		
Genome	Gene ID	Start	Stop	Str.	Start	Extra (nt)	Evaluation
A. radiobacter chr. 2	Arad_7499	430127	431005	fwd.	430130	-3	**Change CTG**
A. radiobacter chr. 2	Arad_7920	817597	818448	fwd.	817543	+54	**Change**
A. radiobacter chr. 2	Arad_8127	988105	989187	fwd.	988120	-15	**Change CTG**
A. radiobacter chr. 2	Arad_8281	1127905	1127375	rev.	1127959	+54	**Change**
A. radiobacter chr. 2	Arad_8316	1158746	1159600	fwd.	1158749	-3	**Change CTG**
A. radiobacter chr. 2	Arad_8341	1183088	1183612	fwd.	1183064	+24	**Change**
A. radiobacter chr. 2	Arad_8351	1190315	1191175	fwd.	1190318	-3	**Change CTG**
A. radiobacter chr. 2	Arad_8840	1594101	1595582	fwd.	1593927	+174	**Change**
A. radiobacter chr. 2	Arad_9098	1816141	1817154	fwd.	1816096	+45	**Change**
A. radiobacter chr. 2	Arad_9333	2023774	2024781	fwd.	2023777	-3	**Change CTG**
A. radiobacter chr. 2	Arad_9399	2087678	2087016	rev.	2087687	+9	**Change CTG**
A. radiobacter chr. 2	Arad_9605	2271724	2270276	rev.	2271709	-15	**Change CTG**
B. anthracis	BA_0265	258100	257393	rev.	258142	+42	**Change**
B. anthracis	BA_5541	5032319	5031801	rev.	5032340	+21	**Change**
B. subtilis	BSU01460	151303	152133	fwd.	151264	+39	**Change**
B. subtilis	BSU25180	2599266	2598616	rev.	2599332	+66	**Change**
B. subtilis	BSU26550	2714140	2713949	rev.	2714209	+69	**Change**
B. subtilis	BSU28870	2953349	2952828	rev.	2953331	-18	**Keep ATT**
Bradyrhizobium BTAi1	BBta_0068	70735	69536	rev.	70759	+24	**Change**
Bradyrhizobium BTAi1	BBta_0178	184778	185194	fwd.	184751	+27	**Change**
Bradyrhizobium BTAi1	BBta_0456	456727	457239	fwd.	456577	+150	**Change**
Bradyrhizobium BTAi1	BBta_0472	470341	470496	fwd.	470215	+126	**Change**
Bradyrhizobium BTAi1	BBta_0740	761034	760111	rev.	761145	+111	**Change**
Bradyrhizobium BTAi1	BBta_0948	981763	982704	fwd.	981736	+27	**Change**
Bradyrhizobium BTAi1	BBta_0951	984964	984527	rev.	985012	+48	**Change**
Bradyrhizobium BTAi1	BBta_1240	1316169	1317290	fwd.	1316007	+162	**Change**
Bradyrhizobium BTAi1	BBta_1457	1543707	1543964	fwd.	1543659	+48	**Change**
Bradyrhizobium BTAi1	BBta_1599	1672857	1672570	rev.	1672893	+36	**Change**
Bradyrhizobium BTAi1	BBta_1707	1780631	1780266	rev.	1780661	+30	**Change**
Bradyrhizobium BTAi1	BBta_1788	1856837	1857526	fwd.	1856831	+6	**Change CTG**
Bradyrhizobium BTAi1	BBta_2009	2074788	2075000	fwd.	2074764	+24	**Change**
Bradyrhizobium BTAi1	BBta_2165	2241836	2243518	fwd.	2241800	+36	**Change**
Bradyrhizobium BTAi1	BBta_2898	3004559	3005311	fwd.	3004529	+30	**Change**
Bradyrhizobium BTAi1	BBta_3040	3177551	3175974	rev.	3177626	+75	**ReviewWeak**
Bradyrhizobium BTAi1	BBta_3463	3614525	3613716	rev.	3614585	+60	**Change**
Bradyrhizobium BTAi1	BBta_3585	3745821	3747008	fwd.	3745767	+54	**Change**
Bradyrhizobium BTAi1	BBta_3602	3767174	3768643	fwd.	3767144	+30	**Change**
Bradyrhizobium BTAi1	BBta_3675	3838804	3840465	fwd.	3838771	+33	**Change**
Bradyrhizobium BTAi1	BBta_3767	3947285	3946362	rev.	3947396	+111	**Change**
Bradyrhizobium BTAi1	BBta_3829	4015302	4012717	rev.	4015368	+66	**Change**
Bradyrhizobium BTAi1	BBta_4168	4379434	4378142	rev.	4379488	+54	**Change**
Bradyrhizobium BTAi1	BBta_4219	4432432	4433520	fwd.	4432330	+102	**ReviewUp**
Bradyrhizobium BTAi1	BBta_4257	4470093	4469278	rev.	4470129	+36	**Change**
Bradyrhizobium BTAi1	BBta_4324	4530792	4531415	fwd.	4530714	+78	**Change**
Bradyrhizobium BTAi1	BBta_4516	4731195	4730293	rev.	4731216	+21	**Change**
Bradyrhizobium BTAi1	BBta_5019	5249017	5249691	fwd.	5248993	+24	**Change**
Bradyrhizobium BTAi1	BBta_5059	5283790	5283476	rev.	5284099	+309	**Change**
Bradyrhizobium BTAi1	BBta_5089	5314187	5314029	rev.	5314220	+33	**Change**
Bradyrhizobium BTAi1	BBta_5101	5326283	5327206	fwd.	5326172	+111	**Change**
Bradyrhizobium BTAi1	BBta_5121	5349491	5348595	rev.	5349536	+45	**Change**
Bradyrhizobium BTAi1	BBta_5178	5407689	5408177	fwd.	5407632	+57	**Change**
Bradyrhizobium BTAi1	BBta_5409	5630107	5629340	rev.	5630137	+30	**ReviewWeak**
Bradyrhizobium BTAi1	BBta_6332	6594995	6595918	fwd.	6594884	+111	**Change**
Bradyrhizobium BTAi1	BBta_6527	6808355	6810403	fwd.	6808175	+180	**Change CTG**
Bradyrhizobium BTAi1	BBta_6829	7153662	7153033	rev.	7153683	+21	**Change**
Bradyrhizobium BTAi1	BBta_6844	7170752	7170844	fwd.	7170638	+114	**Change**

Continued on next page

		RefSeq			PHANTIM		
Genome	Gene ID	Start	Stop	Str.	Start	Extra (nt)	Evaluation
Bradyrhizobium BTAi1	BBta_7012	7346366	7346106	rev.	7346408	+42	**Change**
Bradyrhizobium BTAi1	BBta_7069	7410657	7411580	fwd.	7410546	+111	**Change**
Bradyrhizobium BTAi1	BBta_7206	7566745	7565036	rev.	7566808	+63	**Change**
Bradyrhizobium BTAi1	BBta_7442	7816925	7817194	fwd.	7816703	+222	**Change**
Bradyrhizobium BTAi1	BBta_7457	7833340	7833567	fwd.	7833283	+57	**Change**
Bradyrhizobium BTAi1	BBta_7495	7870110	7869391	rev.	7870164	+54	**Change**
Bradyrhizobium BTAi1	BBta_7569	7953895	7953326	rev.	7953967	+72	**Change**
C. violaceum	CV_0086	97321	99189	fwd.	97279	+42	**Change**
C. violaceum	CV_0099	114821	113448	rev.	115055	+234	**ReviewUp**
C. violaceum	CV_0668	695894	696301	fwd.	695831	+63	**Change**
C. violaceum	CV_0744	764543	764809	fwd.	764453	+90	**Change**
C. violaceum	CV_0766	788326	786851	rev.	788383	+57	**Change**
C. violaceum	CV_0794	818319	817570	rev.	818394	+75	**Change**
C. violaceum	CV_1122	1175575	1176570	fwd.	1175503	+72	**Change**
C. violaceum	CV_1162	1221800	1220469	rev.	1221842	+42	**Change**
C. violaceum	CV_1224	1285455	1287536	fwd.	1285329	+126	**Change**
C. violaceum	CV_1457	1545716	1545402	rev.	1545752	+36	**Change**
C. violaceum	CV_3405	3699237	3699959	fwd.	3699075	+162	**Change**
C. violaceum	CV_3735	4033281	4033877	fwd.	4032915	+366	**ReviewUp**
C. violaceum	CV_3965	4289117	4286490	rev.	4289129	+12	**Change CTG**
C. violaceum	CV_4100	4442560	4441790	rev.	4442584	+24	**ReviewWeak CTG**
C. violaceum	CV_4405	4750493	4750305	rev.	4750514	+21	**Change CTG**
E. coli	b0429	447270	446941	rev.	447351	+81	**Change**
E. coli	b0923	974845	975549	fwd.	974818	+27	**ReviewWeak**
E. coli	b1081	1136594	1137535	fwd.	1136564	+30	**Change**
E. coli	b1432	1501741	1502889	fwd.	1501681	+60	**Change**
E. coli	b1718	1798662	1798120	rev.	1798554	-108	**Keep ATT**
E. coli	b4461	2746796	2748082	fwd.	2746820	-24	**Keep CTG**
H. pylori	HP0885	935407	936792	fwd.	935332	+75	**Change**
H. pylori	HP1507	1580322	1581479	fwd.	1580280	+42	**Change**
M. tuberculosis	MT0055	52775	53188	fwd.	52754	+21	**Change**
M. tuberculosis	MT0436	510004	509207	rev.	510034	+30	**Change**
M. tuberculosis	MT0786	857867	856839	rev.	857966	+99	**Change**
M. tuberculosis	MT1024	1111651	1112262	fwd.	1111609	+42	**ReviewWeak**
M. tuberculosis	MT1041	1131690	1133303	fwd.	1131669	+21	**Change**
M. tuberculosis	MT2188	2390918	2389791	rev.	2391035	+117	**Change**
M. tuberculosis	MT2191	2393215	2392427	rev.	2393320	+105	**ReviewUp**
M. tuberculosis	MT3860	4192534	4192034	rev.	4192555	+21	**Change**
X. bovienii	XBJ1_0065	69975	72482	fwd.	69954	+21	**ReviewUp**
X. bovienii	XBJ1_0068	75246	77303	fwd.	75195	+51	**Change**
X. bovienii	XBJ1_0241	244359	243160	rev.	244416	+57	**Change**
X. bovienii	XBJ1_0338	346775	345549	rev.	346880	+105	**Change**
X. bovienii	XBJ1_0342	349583	353326	fwd.	349559	+24	**Change**
X. bovienii	XBJ1_0604	621716	625150	fwd.	621668	+48	**Change**
X. bovienii	XBJ1_0731	718904	721669	fwd.	718835	+69	**Change**
X. bovienii	XBJ1_0803	822940	822218	rev.	822967	+27	**ReviewWeak**
X. bovienii	XBJ1_0893	903055	904665	fwd.	902896	+159	**Change**
X. bovienii	XBJ1_1116	1139112	1140110	fwd.	1139070	+42	**Change**
X. bovienii	XBJ1_1800	1758537	1759787	fwd.	1758480	+57	**Change**
X. bovienii	XBJ1_1924	1859695	1861266	fwd.	1859635	+60	**Change**
X. bovienii	XBJ1_1926	1862240	1863208	fwd.	1862156	+84	**Change**
X. bovienii	XBJ1_1960	1892419	1893384	fwd.	1892389	+30	**ReviewWeak**
X. bovienii	XBJ1_2624	2590238	2590903	fwd.	2590196	+42	**Change**
X. bovienii	XBJ1_2675	2630754	2628988	rev.	2630820	+66	**Change**
X. bovienii	XBJ1_2786	2754576	2753962	rev.	2754792	+216	**ReviewUp**
X. bovienii	XBJ1_2841	2792338	2791175	rev.	2792377	+39	**Change**

Continued on next page

		RefSeq			PHANTIM		
Genome	Gene ID	Start	Stop	Str.	Start	Extra (nt)	Evaluation
X. bovienii	XBJ1_2847	2800875	2800045	rev.	2800932	+57	**ReviewWeak**
X. bovienii	XBJ1_2862	2811808	2810801	rev.	2811838	+30	**Change**
X. bovienii	XBJ1_2952	2918549	2919121	fwd.	2918501	+48	**Change**
X. bovienii	XBJ1_3332	3256054	3256938	fwd.	3256009	+45	**Change**
X. bovienii	XBJ1_3473	3382791	3382066	rev.	3382890	+99	**Change**
X. bovienii	XBJ1_3487	3398596	3398105	rev.	3398626	+30	**ReviewWeak**
X. bovienii	XBJ1_3802	3682635	3683657	fwd.	3682593	+42	**ReviewWeak**
X. bovienii	XBJ1_3913	3789753	3788374	rev.	3789783	+30	**Change**
X. bovienii	XBJ1_4158	3995207	3992748	rev.	3995348	+141	**Change**
X. bovienii	XBJ1_4169	4006207	4005245	rev.	4006324	+117	**Change**
X. bovienii	XBJ1_4412	4224742	4224416	rev.	4224775	+33	**Change**