GenePRIMP: A GENE PRediction IMprovement Pipeline for Prokaryotic genomes

Amrita Pati[1], Natalia N. Ivanova[1], Natalia Mikhailova[1], Galina Ovchinnikova[1], Sean D. Hooper[1,2], Athanasios Lykidis[1] & Nikos C. Kyrpides[1]

[1] Genome Biology Program, Joint Genome Institute, 2800 Mitchell Dr, Walnut Creek, CA, USA. Correspondence should be addressed to A.P. (apati@lbl.gov).
[2] Present address: Department of Genetics and Pathology, Uppsala University, SE-751 85 Uppsala, Sweden.

We present GenePRIMP (Gene Prediction IMprovement Pipeline, http://geneprimp.jgi-psf.org), a computational process that performs evidence-based evaluation of gene models in prokaryotic genomes and reports anomalies including inconsistent start sites, missed genes, and split genes. We show that manual curation of gene models using the anomaly reports generated by GenePRIMP improves their quality and demonstrate the applicability of GenePRIMP in improving finishing quality and comparing different genome sequencing and annotation technologies.

More than 1000 microbial genomes have been completely sequenced to date[1]. The increasing number of sequencing projects driven by high-throughput sequencing technologies has further underscored the importance of computational methods in annotating and mining genomic data. For any genome, gene finding is the key step to understanding the biochemistry, physiology, and ecology of the organism. Gene finding relies heavily on computational methods and very few sequencing projects are complemented by the experimental verification of computationally predicted genes through functional genomics experiments or mapping of N-terminal sequences[2,3]. Together with multiple sequencing technologies, multiple gene finders, and somewhat imprecise standards for the identification of genes, this can result in different researchers arriving at substantially varying gene models for the same organism[4] (**Fig. 1**, **Table 1**). Consequently, higher standards of accuracy are required for computational gene prediction tools.

The most popular gene finders are *ab initio* and work by statistically profiling protein coding, intergenic, and boundary regions using a variety of classifiers. While most *ab initio* gene callers boast an average accuracy of 90% or better[5-7], accuracy can be compromised by many factors such as genomic islands of differing GC content, pseudogenes, and genes with programmed or artificial frameshifts, leading to sizeable variability between their gene model predictions. To improve gene models generated by *ab initio* predictions, some tools include heuristics and post-processing steps such as overlap removal, translation initiation site adjustment, and frameshift detection[8,9], while others rely on the presence of sequenced close relatives[10] or experimental evidence[11,12]. However, many of these post-processing tools have been tested only on metazoan genomes and use criteria that are not applicable to prokaryotes, and/or are too slow or expensive to perform on a large number of microbial genomes.

To overcome the aforesaid limitations of *ab initio* gene prediction methods, and to address the problem of large variation among their gene models, we have devised GenePRIMP; a computational evidence-based post-processing pipeline that identifies erroneously predicted

genes. Manual correction of GenePRIMP-reported genes results in a standardized output gene complement for an organism (sequence) irrespective of the method used for initial gene predictions (**Fig. 1**) [to what extent is manual correction needed in the GenePRIMP pipeline and how can you ensure that this manual correction will be standardized – do you mean that the corrections found with GenePRIMP will then have to be manually added to the list people are working with? The GenePRIMP report only contains the list of problems in gene definitions. These problems need to be corrected manually by the curator. Working with the GenePRIMP-report as a guide, everyone will make more or less the same corrections.]. Other applications of GenePRIMP include benchmarking of *ab initio* gene callers, improvement of finishing quality, detection of frameshifts in sequences generated by various technologies, and application to fungal and eukaryotic genomes with minor changes in the associated heuristics (Methods).

A typical GenePRIMP report includes seven types of anomalies, namely: short genes, long genes, unique genes, dubious genes, broken genes, interrupted genes, and putative missed genes (**Fig. 2**), identified from the alignment of a gene or intergenic region to its homologs. While short and long genes have anomalous start sites, broken and interrupted genes are parts of the same gene called as multiple genes. Unique/dubious genes, which have no hits to known proteins, may reveal a perfectly good gene in a different frame; such hits are included in the list of putative missed genes when examined together with the bounding intergenic regions by BLASTx. Alternatively, they may be experimentally verifiable novel genes. Broken genes might indicate the presence of a pseudogene, a programmed frameshift that does not render the gene non-functional, or a frameshift due to sequencing artifacts (for example base calling errors in homopolymer regions). GenePRIMP ensures that fusion gene components are not mislabeled as frameshift-induced broken genes by comparing against a database of fusion genes[13]. Joining of said frameshift fragments and subsequent tagging of genes is at the sole discretion of the curator.

The protocols captured in GenePRIMP are a result of the standardization of operating procedures used in the DOE-JGI in the manual curation of over 300 genomes (>100,000 genes), coming from multiple sequencing centers, over a period of 3 years. Over 194 genomes (>400 contigs, including permanent drafts) have been processed by GenePRIMP followed by manual curation (**Supplementary Data 1**). On average, about 10% of the genes in a given genome are modified by manual curation, but this percentage varies between 3% and 20% depending on the properties of the genome and the gene finder software used. With the current version of GenePRIMP, approximately 85% of all reported short genes are manually extended (short genes can only be extended with evidence when there is space on the 5' end for extension), 70% of all reported long genes are manually truncated, and 100% of reported broken genes as well as 31% of reported interrupted genes are manually joined. We have not shown statistics for putative missed genes because some of these intergenic regions with hits are combined with short genes during extension. We find that the numbers of short, long, unique, and total reported anomalies are positively correlated with genome size (R^2 = 0.66, 0.65, 0.38, 0.407, respectively), but no correlation of anomalies is observed with genome GC content (R^2 = 0.0007, 0.1038, 0.0004, 0.0134, 0.0076, 0.00006, 0.0023 for short, long, unique, broken, interrupted, missed genes, and total number of reported anomalies, respectively). We observe positive correlations between some anomaly types: a moderately high correlation between the numbers of short and long genes, likely arising from imprecise detection of ribosome-binding sites by *ab initio* gene finders (**Supplementary Fig. 1).**

We used GenePRIMP to compare the accuracy of five popular gene finders: Prodigal[5], GeneMark[6], Glimmer3[7], RAST[14], and AMIGene[15] by evaluating their gene calls for two

genomes: the bacterium *Mycobacterium* sp. Spyr1 (Myco) and the archaeon *Methanosphaerula palustris* E1-9c (Meth), selected because of the high number of modifications made to their gene models during manual curation (See **Supplementary Data 2-5** for gene definitions in these two genomes before and after manual curation). Comparisons were based on the number of anomalies of each type detected by GenePRIMP (**Table 1**). Results of automated gene finding for these two genomes vary wildly among the different tools and pipelines. Notably in Meth, Glimmer3 predicted the most unique genes (522); 226 of these were not called by any other gene caller and only 38 genes were predicted by all others (**Supplementary Fig. 2**). Glimmer3 identified 515 more genes (18%) in Meth than did Prodigal, which identified the lowest number of genes; many of these additional genes were among the 522 unique genes predicted by Glimmer3. We observed considerable variation in the gene-finders' identification of translation initiation sites. Glimmer3, GeneMark, and RAST show a tendency to predict genes shorter than their homologs, whereas AMIGene calls more long genes than any of the others. The occurrence of missed genes and predicted genes that are longer or shorter than their homologs reflects the current limitations of automated gene finding in microbial genomes. The number of broken and interrupted genes identified in the gene calls indicates the sensitivity of the respective gene caller. Higher numbers attest to the greater ability of that gene caller to identify shorter regions of CDSs, including small fragments in highly degraded pseudogenes. This facilitates the correction of sequencing artifacts, pseudogenes, and genes with unusual translational features.

After manual curation for a given genome using the GenePRIMP report, the final gene model complements are very similar (**Fig. 1**) even though different gene callers are used. Data on genes from Meth, with no closely related sequenced genomes, demonstrate that the accuracy of GenePRIMP does not rely heavily on the presence of closely related species. We examined the 2584 genes that have matching stop positions and differing start positions, and the 1669 genes that have both matching start positions and matching stop positions, among gene calls of three gene callers: Prodigal, GeneMark, and RAST. These are further examined in Supplementary Fig. 3. [is it necessary to detail these results in the main text – you could just refer to SI fig. 3. I have changed the text accordingly.] Since Prodigal is part of the regular microbial annotation pipeline at the DOE-JGI, we evaluated GenePRIMP's handling of the 235 genes that were only called by Prodigal (8% of the Prodigal total). Hypothesizing that most of those 235 genes are good predictions, we deduced that GenePRIMP should report them as missed genes in the RAST and GeneMark gene calls. From **Fig. 1**, we observe that GenePRIMP correctly identified 93% of the 235 genes predicted by Prodigal but missed by RAST or GeneMark or both. Examination of the remaining 7% revealed that GenePRIMP did not discover them because of the presence of spurious genes on the opposite strands.

GenePRIMP is available as a web-based application (**Supplementary Fig. 4**). The compute time for any genome is dominated by the time taken to perform Blast alignments; a 4 Mb genome typically runs in about 2 hours on a computer with 16 2300 MHz CPUs and 64 GB of shared memory. Current and future directions for GenePRIMP include automatic correction of GenePRIMP-reported anomalies to the extent possible, as well as automatic identification of putative frameshifts and pseudogenes. GenePRIMP is a significant step towards automation and standardization of the long-standing process of gene finding and manual curation. As such, it is also following the principles of standardization of the Genomics Standards Consortium and further development will factor in the Consortium's recommendations.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

Acknowledgements
We would like to gratefully acknowledge the help and support of I. Anderson and K. Mavromatis from the Genome Biology Program, Xueling Zhao from DOE JGI, and V. Markowitz from the Biological Data Management and Technology Center. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

AUTHOR CONTRIBUTIONS
N.N.I. and N.C.K. conceived the initial approach. N.N.I. and A.P. designed the system. A.P. implemented the GenePRIMP code base and web portal. S.D.H. contributed to the development of the web portal. N.N.I., N.M., G.O., and A.L. did the manual curation for the genomes and contributed to testing and validation.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

References

1. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. *Nucleic Acids Res*. **38**, D46-D51 (2010).
2. Ishino, Y., Okada, H., Ikeuchi, M. & Taniguchi, H. *Proteomics*. **7**, 4053-4065 (2007).
3. Smollett, K.L. *et al*. *Microbiology* **155**, 186-197 (2009).
4. Kyrpides, N.C. *Nat. Biotechnology* **27**, 627 -632 (2009)
5. Hyatt, Doug. *et al*. *BMC Bioinformatics*, In press.
6. Besemer, J., Lomsadze, A. & Borodovsky, M. *Nucleic Acids Res*. **29**, 2607-2618 (2001).
7. Delcher, A.L., Bratke, K.A., Powers, E.C. & Salzberg, S.L. *Bioinformatics* **23**, 673-679 (2007).
8. Zhu, H.Q., Hu, G.Q., Quyang, Z.Q., Wang, J. & She, Z.S. *Bioinformatics* **20**, 3308-3317 (2008).
9. Tech, M. & Meinicke, P. *BMC Bioinformatics* **7**:121, (2006).
10. Yu, G.X. *et al*. *Nucleic Acids Res*. **35**, 3953-3962 (2007).
11. Nagy, A. *et al*. *BMC Bioinformatics* **9**:353 (2008).
12. Castellanaa, N.E. *et al*. *PNAS* **105**, 21034-21038 (2008).
13. Markowitz V.M., *et al*. *Nucleic Acids Res*, **38**, D382-D390 (2010).
14. Aziz, R.K. *et al*. *BMC Genomics* **9**:75 (2008).
15. Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G., & Medigue, C. *Nucleic Acids Res*. **31**, 3723-3726 (2003).

FIGURE LEGENDS

Figure 1 GenePRIMP analysis of gene calls in *Methanosphaerula palustris* E1-9c (Meth) by three gene callers.. Using GenePRIMP, we analyzed the 2,819-2,584=235 genes in Prodigal that were not common to all of the three gene callers. Of the 121 genes predicted only by Prodigal and GeneMark, GenePRIMP reported 118 as putative missed genes among the RAST gene calls. Likewise, of the 26 genes called only by Prodigal and RAST, GenePRIMP identified 23 as putative missed genes among the GeneMark gene calls. Lastly, of the 88 genes called only by Prodigal, GenePRIMP identified 76 as putative missed genes in both the GeneMark and RAST gene calls. 7 of these 88 genes were found in neither GeneMark- nor RAST-generated gene calls. For 83 of the 88 genes (corresponding cells are highlighted in yellow), GenePRIMP decisions matched for both GeneMark- and RAST-generated gene calls. For only 5 of these 88 genes, GenePRIMP decisions disagreed between GeneMark- and RAST-generated gene calls. The disagreement was due to the presence of spurious genes on the opposite strands in the same intergenic region for one of the two gene callers. Similar results were observed for Glimmer3 and Amigene (data not shown). Please do not repeat the numbers already stated in the figure, but explain what the yellow color in the 3rd column means. Since we have pruned the descriptions in the text so much, the existing figure legends are necessary for the explanations to flow.

Figure 2 The GenePRIMP processing pipeline. (a) Detection of gene call anomalies by GenePRIMP. (b) Blast alignments of short, long, broken, and interrupted genes. A query gene is shown aligned against its homologs in NCBI's nr database for each of the indicated classes. All sequences are shown 5' to 3' from left to right. [this is already explained in the main text. The exact definitions of broken and interrupted genes are not included in the main text. Also, short and long genes are explained in greater detail in this legend. Since these descriptions are key to understanding the anomalies, I would prefer to keep this part.]

Table 1 Comparison of five gene calling applications.

	Mycobacterium sp. Spyr1 GC% = 67.9, Size=6 Mb					Methanosphaerula palustris E1-9c GC% = 55.35, Size=2.9 Mb				
	GeneMark	Glimmer3	Prodigal	RAST	AMI Gene	Gene Mark	Glimmer3	Prodigal	RAST	AMI Gene
CDSs	5553	5395	5296	5304	4888	2974	3334	2819	2940	3177
Short genes	482	398	267	672	79	235	230	202	420	115
Long genes	83	53	62	34	992	46	59	60	47	294
% CDSs w/ anomalous starts (short + long)	10.17	8.36	6.21	13.31	21.74	9.49	8.69	9.61	15.9	12.87
Missed genes	607 (10.93 %)	569 (10.54 %)	451 (8.51 %)	735 (13.9 %)	658 (13.46 %)	196 (6.59 %)	206 (6.18%)	167 (5.92%)	305 (10.4%)	106 (3.33 %)
Unique genes	67	118	23	206	99	190	522	103	229	277
Dubious genes	11	0	2	0	10	25	0	2	1	8
Broken genes	30	33	27	22	34	41	50	27	29	71
Interrupted genes	51	62	48	60	53	23	36	32	31	60

Online Methods
All text needs to be run on and only one level of subheadings are allowed. Done.

The GenePRIMP algorithm

The flowchart for the high-level GenePRIMP algorithm is diagrammed in **Supplementary Fig. 5**. In summary, for each contig in the input file, all features are parsed and stored. PILER-CR, a CRISPR finder, is then run on the contig sequence and any CRISPRs found are integrated into the feature list. Any overlaps between features are computed and an overlap report generated. Protein sequences for genes are aligned to a low-complexity filtered Blast database using the parameters '-p blastp -e 0.00001 -b 15 -v 15 -a 16'. Genes without hits are aligned again using Blast with a relaxed cutoff with parameters '-p blastp -e 10 -b 15 -v 15 -F F -a 16'. Genes without hits are classified as unique. Unique genes that are shorter than 30 amino acids are classified as dubious. Genes with hits from both rounds of Blast are filtered to removeintersecting high-scoring pairs (HSPs) constituting bad alignments, hits to eukaryotes, and hits to the draft genome of the same subject organism. Filtered alignments are used for the classification into long/short/broken/interrupted and potential long/short genes. The exact algorithms for identifying long, short, broken, and interrupted genes are described below. Intergenic regions are computed that include unique and dubious genes and boundary adjustments for short and long genes. These adjusted intergenic regions are aligned to the filtered Blast database with the parameters '-p blastx -e 0.1 -b 10 -v 10 -w 15 -a 16'. Alignments are filtered to remove hits to eukaryotes, HSPs in different frames, and hits to the draft genome of the same subject organism. Intergenic regions with reliable alignments are reported as putative missed genes. Genes that were classified as potential long/short genes are examined further. If a potential short gene and its 5' intergenic region share hits to common subject(s), the gene is confirmed as short. If a potential long gene has a promoter region that is shorter than 100 bases, it is confirmed as long.

Detection of short and long genes

Short and long genes are detected using a criterion called an alignment score (**α**). Let S_Q be a query sequence aligned against homologous sequence S_H. Let c_q and c_h indicate the start coordinates of the alignment on S_Q and S_H, respectively. The alignment quality score (α) is then defined as:

$$\alpha = \frac{c_q - c_h}{c_q + c_h}.$$

While the difference between the start sites, c_q-c_h, is necessary to determine whether a query gene might be long, short, or good, it is not sufficient. **Supplementary Figs. 6**(**a**) and **6**(**b**) illustrate two candidates for long genes. Observe that for the same difference in starts of alignment (c_q-c_h =28), whether a gene is long or not also depends on where the alignment starts on the subject and the query. The same phenomenon for short genes is illustrated in **Supplementary Figs. 6**(**c**) and **6**(**d**). Therefore, instead of making decisions based simply on the difference between the start sites, c_q-c_h, GenePRIMP uses the alignment quality score that represents the disparity in the start positions as a fraction of the actual distance of the start

positions from the beginning of their respective sequences (**Supplementary Fig. 7**). To obtain the cut-off values shown, we plotted the distribution of mean and median values of α for genes from five genomes that had been manually curated and identified as long, short, or matching genes (**Supplementary Figs. 8** and **9**). The resultant mean and median alignment scores were also calculated for each gene type (**Supplementary Table 1**).

Detection of broken genes

Two genes that are called adjacent to one another are identified as a broken gene if they satisfy all the following conditions: they have the same orientation; they have at least two common homologs; their hits are to consecutive regions on the same homolog, not to the same regions; their shared homologs are approximately of the same length; when the sequence from the beginning of the first gene to the end of the second gene is aligned using BlastX, at least one hit is among those observed for the two individual genes; the hits for the combined region are not fusion genes as recognized by the fusion genes database in IMG.

Detection of interrupted genes

Two genes are identified as a gene interrupted by a transposase(s) when all the following conditions are satisfied: they have the same orientation; they have at least one common homolog; each of them has hits to at least 4 subjects; the homologs are approximately of the same length.

Application to fungal and eukaryotic genomes

With minor changes, GenePRIMP can be applied to fungal/eukaryotic genomes. Hits from both BlastP and BlastX for genes and intergenic regions, respectively, are filtered to exclude hits to eukaryotic genomes. Additionally, these filters as well as other filters in various stages of the GenePRIMP algorithm employ various heuristics based on the number of hits to eukaryotic genomes present. Adjustment of these filters and heuristics leads to an anomaly-detection framework for eukaryotic and fungal genomes.

Detection of frameshifts

Putative frameshifts can be detected from the following scenarios of anomalies: when a gene is short on the 5' end or on the 3' end and the remaining fragment of the gene is present in the adjoining intergenic region in a frame different from that of the gene; when a missed gene is inserted with one or more frameshifts; when two or more genes are determined as part of the broken gene and are joined with one or more frameshifts; convergent overlaps.

These methods of frameshifts detection have been implemented in an automated manner in the next version of GenePRIMP and are currently being tested in-house. With the current version of GenePRIMP, these scenarios have to be manually analyzed to detect frameshifts.

Detection of pseudogenes

As with frameshifts, pseudogenes can be detected by analyzing specific anomaly scenarios. The following anomaly scenarios suggest the presence of candidate pseudogenes: 1. A gene short

either on the 5’ end or the 3’ end, which cannot be extended either due to the absence of sufficient intergenic space on the said side, or due to the absence of the correct sequence in the available intergenic region. Such a gene might miss important functional domains and become non-functional. While the “pseudo” status for such a gene can be manually determined with relatively high confidence, the automatic detection of “pseudo” status for such a gene might be possible through the analysis of its alignment to homologs. 2. A short gene that has been extended on either the 3’ or the 5’ end, or a missed gene that has been inserted with one or more frameshifts. In the case of two or more frameshifts, loss of function is definite and the gene can be automatically tagged as a pseudogene. In the case where only one frameshift is present, it might be because of a ribosomal slippage. This scenario requires manual analysis to confirm pseudo status. 3. A short gene that has been extended on either the 3’ or the 5’ end, or a missed gene that has been inserted with one or more stop codons or multiple frameshifts or stop codons. The presence of multiple frameshifts/stop codons is once again a positive indicator for a pseudogene. 4. A broken gene that is joined with multiple frameshifts/stop codons can also indicate a putative pseudogene.

These methods of pseudogene detection have been implemented in an automated manner in the next version of GenePRIMP and are currently being tested in-house. With the current version of GenePRIMP, these scenarios have to be manually analyzed to detect pseudogenes.

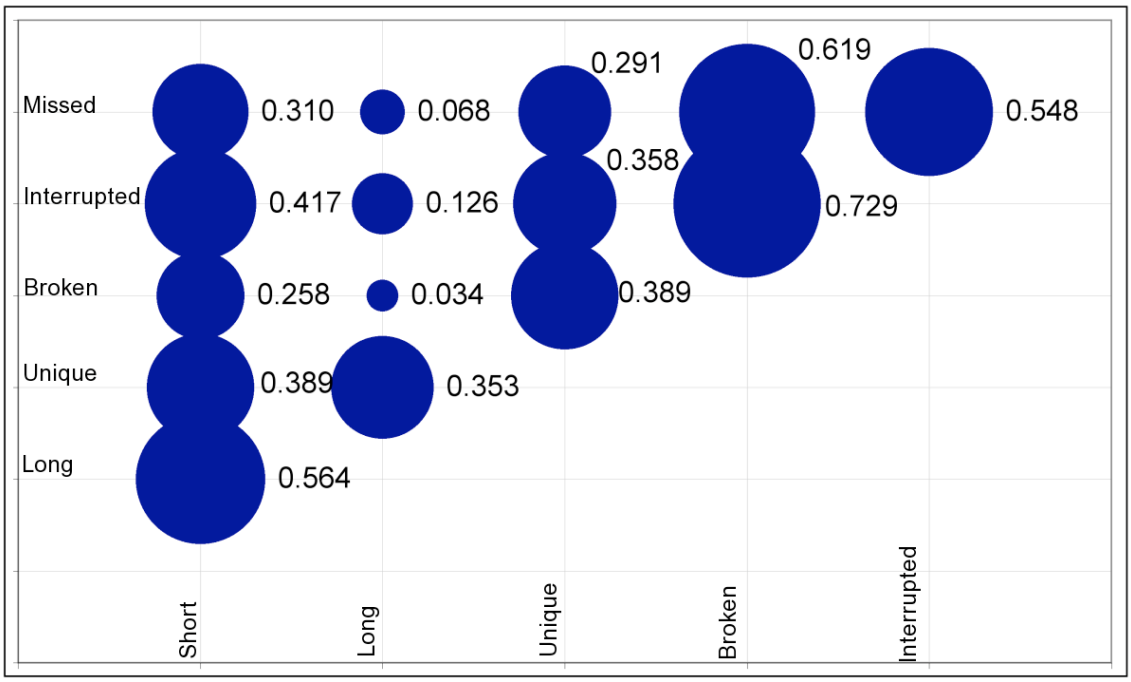
LIST OF SUPPLEMENTARY ITEMS

Supplementary File	Title
Supplementary Figure 1	Correlations between anomaly types.
Supplementary Figure 2	Analysis of unique genes identified by Glimmer3 and RAST.
Supplementary Figure 3	Comparison of GenePRIMP anomaly reports with long/short/unique gene calls.
Supplementary Figure 4	Collage showing screenshots of web pages in the GenePRIMP portal.
Supplementary Figure 5	Process flow in GenePRIMP.
Supplementary Figure 6	Identification of long and short genes from alignments.
Supplementary Figure 7	Identification of long, short, and matching genes based on average and median alignment scores (α).
Supplementary Figure 8	Distribution of mean alignment scores (α) in short, matching, and long genes.
Supplementary Figure 9	Distribution of median alignment scores (α) in short, matching, and long genes
Supplementary Table 1	Alignment scores (α) for short, matching, and long genes.
Supplementary Data 1	Statistics of all public contigs processed by GenePRIMP
Supplementary Data 2	Gene models in Methanosphaerula palustris E1-9c before manual curation
Supplementary Data 3	Gene models in Methanosphaerula palustris E1-9c after manual curation
Supplementary Data 4	Gene models in Mycobacterium sp. Spyr1 before manual curation
Supplementary Data 5	Gene models in Mycobacterium sp. Spyr1 after manual curation

Editorial Summary

This computational process evaluates gene models in prokaryotic genomes, independent of the gene finder used, and reports anomalies that can be used to improve the quality of gene models through manual curation.

Supplementary Figure 1: Correlations between anomaly types



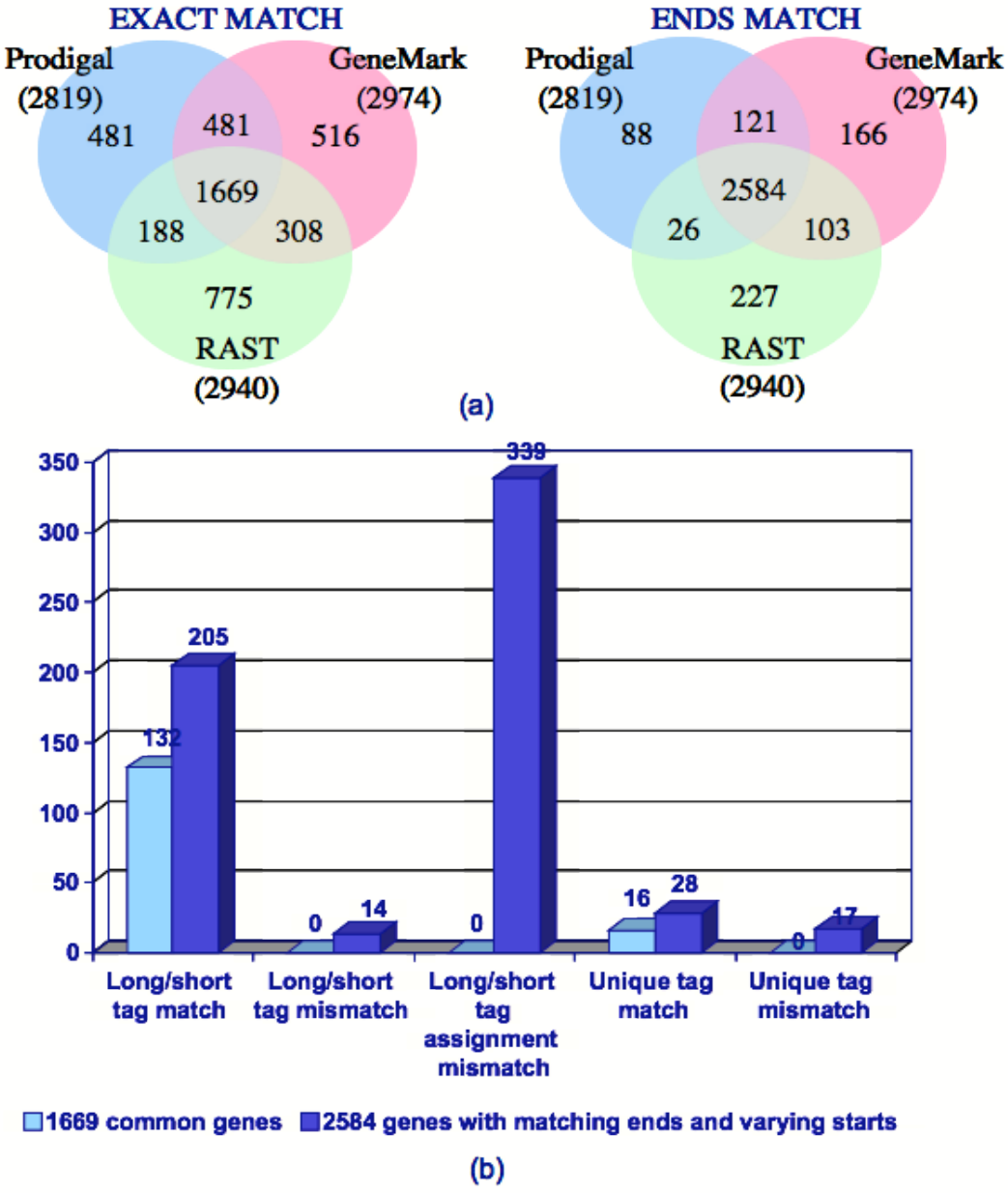
Correlations between anomaly types. Based on all contigs that were manually curated, some types of anomalies were found to be positively correlated with others. The number of missed genes shows the strongest positive correlation with the number of broken, interrupted and short genes, since BLASTx searches of the intergenic regions identified not only missed genes per se, but also missed fragments of predicted genes, such as missed N-terminal sequences of short genes or missed fragments of broken genes. On the other hand, since unique genes often mask a missed gene in a different translation frame, positive correlation between the number of missed and unique genes is not entirely unexpected. Positive correlation between the number of short, interrupted and broken genes likely reflects the fact that these 3 categories include a large number of pseudogenes that may have been generated by gene truncation, disruption of translation frame by frameshifts and/or stop codons or interruption of genes by transposable elements.

Supplementary Figure 2: Analysis of unique genes identified by Glimmer3 and RAST

GeneMark	RAST in case of Meth, Glimmer3 in case of Myco	Prodigal	AMIGene	Unique genes predicted by Glimmer3 in Meth	Unique genes predicted by RAST in Myco
				38	226
				3	62
				0	5
				0	2
				104	56
				57	22
				1	5
				0	5
				0	23
				0	11
				0	4
				0	13
				0	18
				0	22
				3	10
				0	38
Total:				206	522

Analysis of unique genes identified by Glimmer3 and RAST. While Glimmer3 identified the most unique genes in Methanosphaerula palustris E1-9c (Meth), RAST identified the most unique genes in Mycobacterium sp. Spyr1 (Myco). We compared these unique genes to evaluate how many were common with genes called by other gene callers for the same genomes. Red indicates absence and green indicates presence of a gene in the gene calls for a given gene caller. Glimmer3 called the highest number of unique genes in Meth (522) (See Table 1). However, only 38 of these were called by the other 4 gene callers as well. Most of the unique genes (226) were not called by any other gene caller. RAST called the highest number of unique genes in Myco (206). None of these genes were called by all 4 remaining gene callers and 38 out of the 206 genes were not predicted by any other gene caller. 104 of these 206 unique genes were predicted by Glimmer3 as well. This could be explained by the fact that Glimmer3 is part of the RAST pipeline for gene prediction.

Supplementary Figure 3: Comparison of GenePRIMP anomaly reports with long/short/unique gene calls



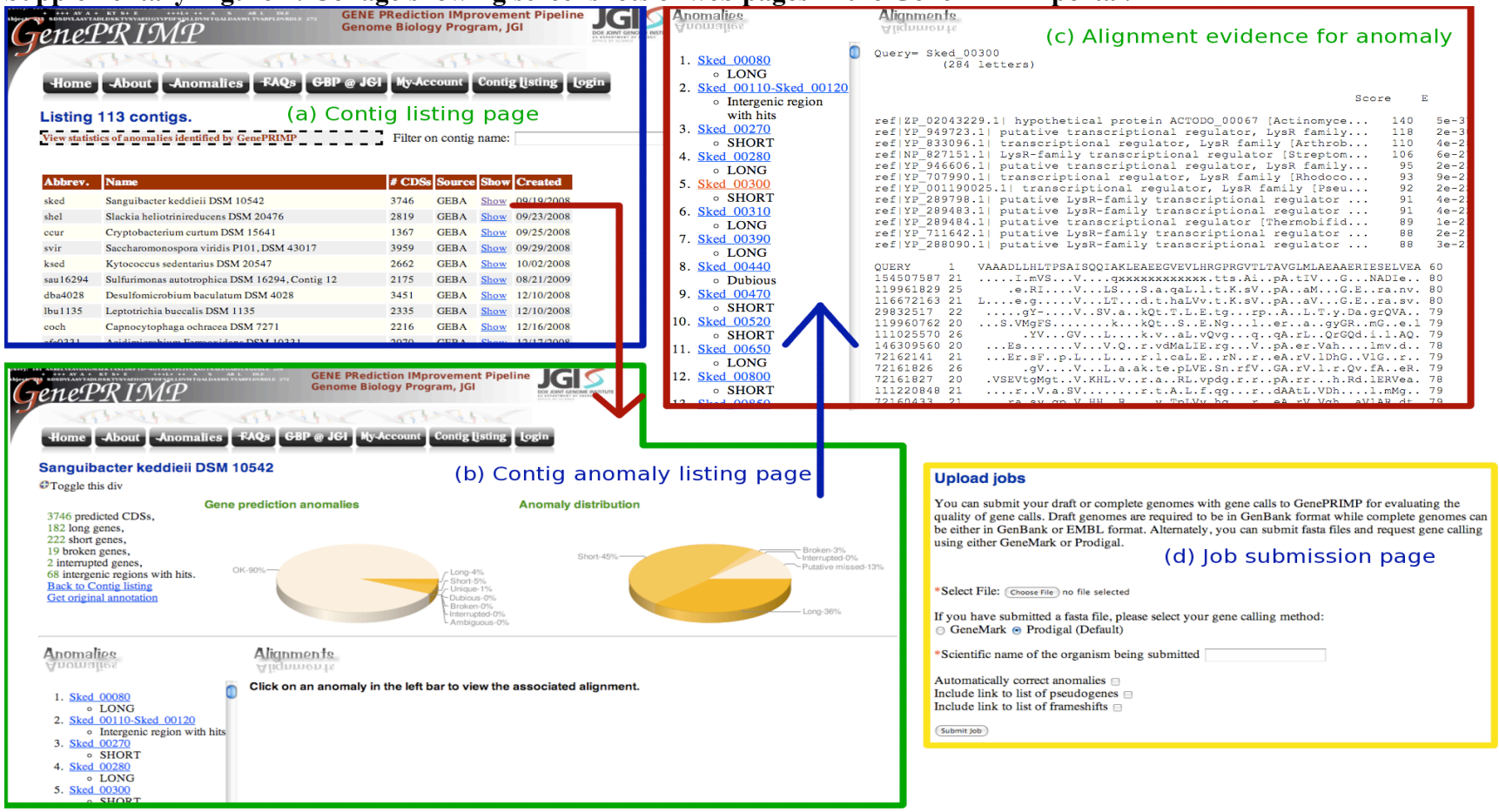
Comparison of GenePRIMP anomaly reports with long/short/unique gene calls. Gene calls generated for the archaeon Methanosphaerula palustris E1-9c by Prodigal, GeneMark, RAST, Glimmer3, and AMIGene were compared and also processed with GenePRIMP. Fig. 3(a) shows comparison of gene models predicted by Prodigal, GeneMark and RAST including comparison of exactly matching genes (EXACT MATCH, i.e., the genes that were predicted on the same strand with identical coordinates of start and stop codons) and of the genes with matching stop codons (ENDS MATCH, i.e., the genes that were predicted on the same strand with identical coordinates of the stop codon, but possibly different start codons). This figure shows that less than 60% of the genes predicted by any single gene caller are exactly matched by both other gene finders, whereas the number of the genes sharing the same stop codon is close to 90%. Similar results

were observed for Glimmer3 and AMIGene as well (not shown for clarity of presentation). Fig. 3(b) shows distribution of anomalies among the genes predicted identically by the 3 gene finders (Prodigal, GeneMark and RAST) and among the genes sharing the same stop codon.

- Unique tag match = genes called as unique by one gene caller and called the same by all three.
- Unique tag mismatch = genes called as unique by one gene caller but not called the same by all three.
- Long/short tag match = genes called as either long or short by one gene caller and called the same by all three.
- Long/short tag/assignment mismatch = genes called as either long or short by one gene caller but not called the same by all three.

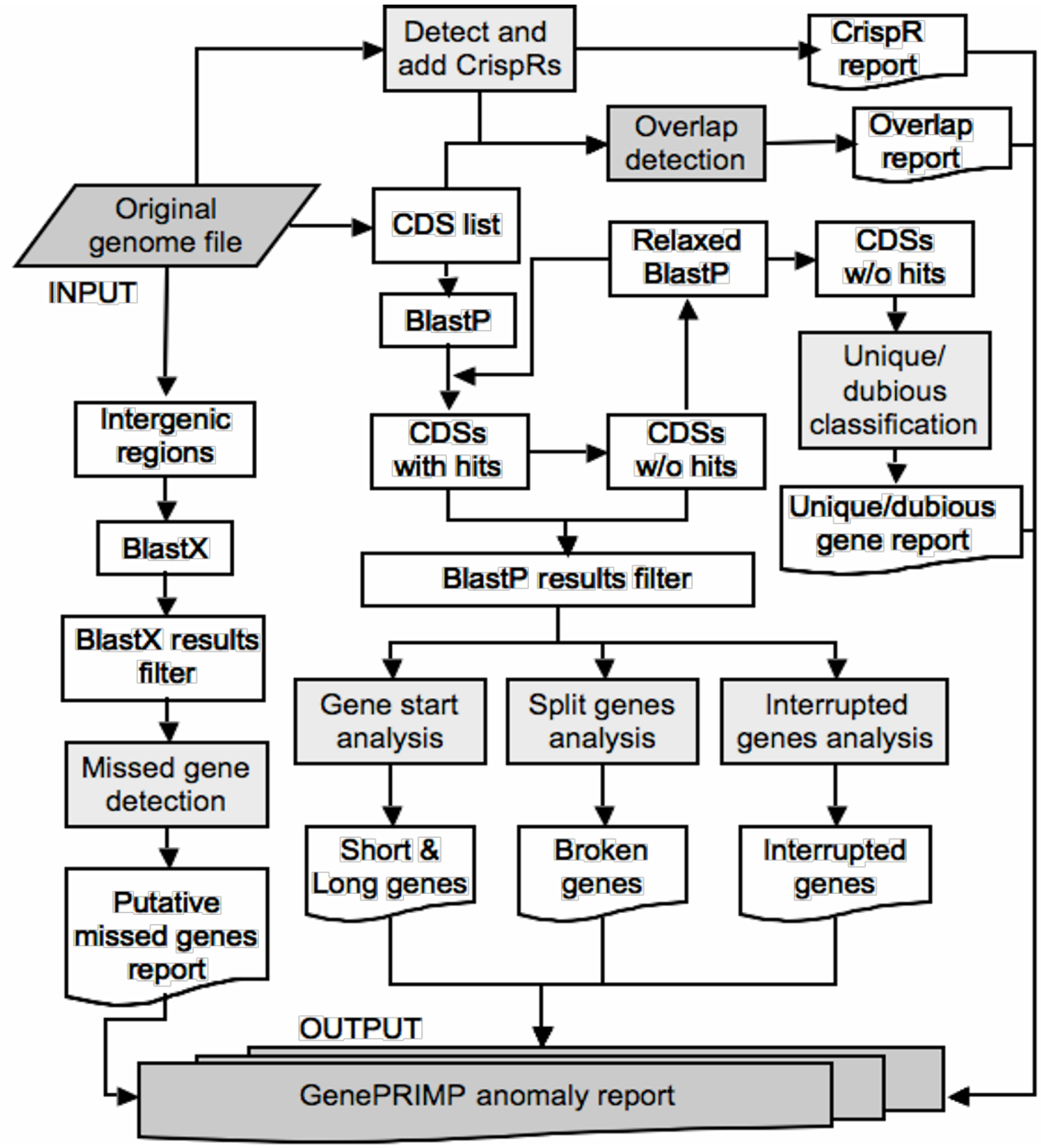
This sanity check shows that if a gene has been called short/long with an anomalous translation start site, GenePRIMP captures it as an anomaly irrespective of the gene caller whose gene definitions are the source of the anomaly.

Supplementary Figure 4: Collage showing screenshots of web pages in the GenePRIMP portal.



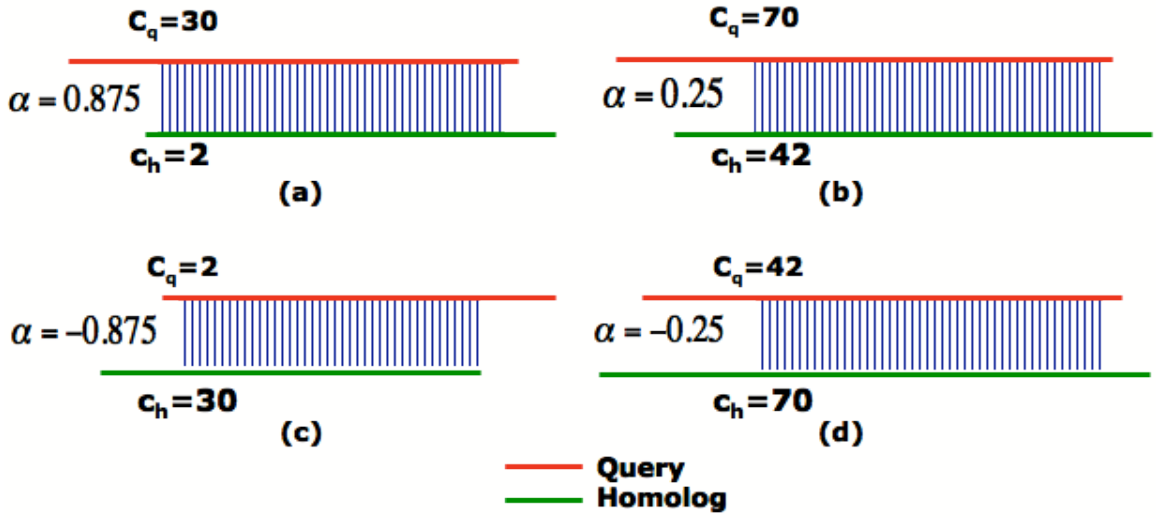
Collage showing screenshots of web pages in the GenePRIMP portal. Evaluated contigs are listed on the contig-listing page (a). From here, statistics and details of the anomalies found for any contig can be viewed by clicking on the “Show” link for that contig and navigating to the contig anomaly-listing page (b). Alignment evidence for each anomaly can be seen by clicking on that anomaly (c). GenePRIMP offers the option to register and submit contigs for processing at the web site.

Supplementary Figure 5. Process flow in GenePRIMP



Process flow in GenePRIMP.

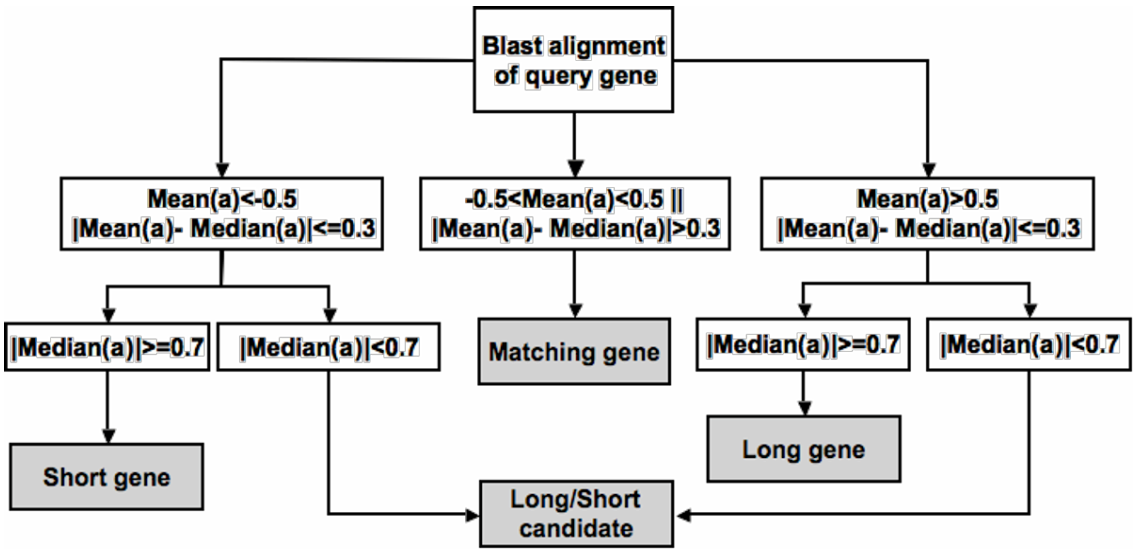
Supplementary Figure 6: Identification of long and short genes from alignments



Identification of long and short genes from alignments.

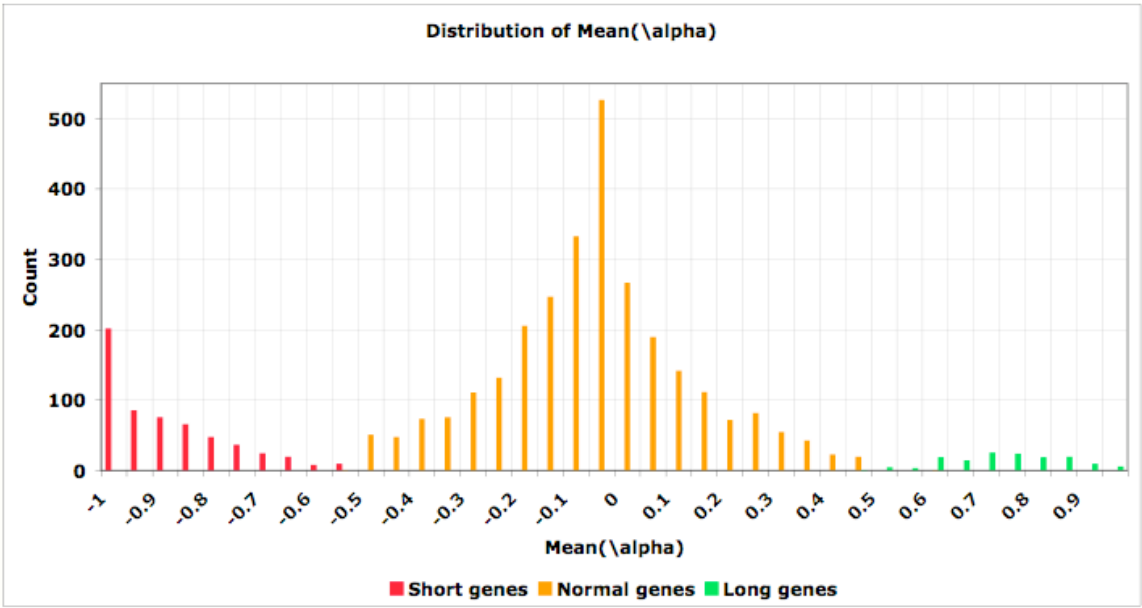
(a) Query gene is actually longer than the subject gene. (b) Query gene appears to be longer than the subject gene as seen from c_q-c_h but is not very much longer. (c) Query gene is actually shorter than the subject gene. (d) Query gene appears to be shorter than the subject gene from c_q-c_h but is not very much shorter.

Supplementary Figure 7: Identification of long, short, and matching genes based on average and median alignment scores (α)



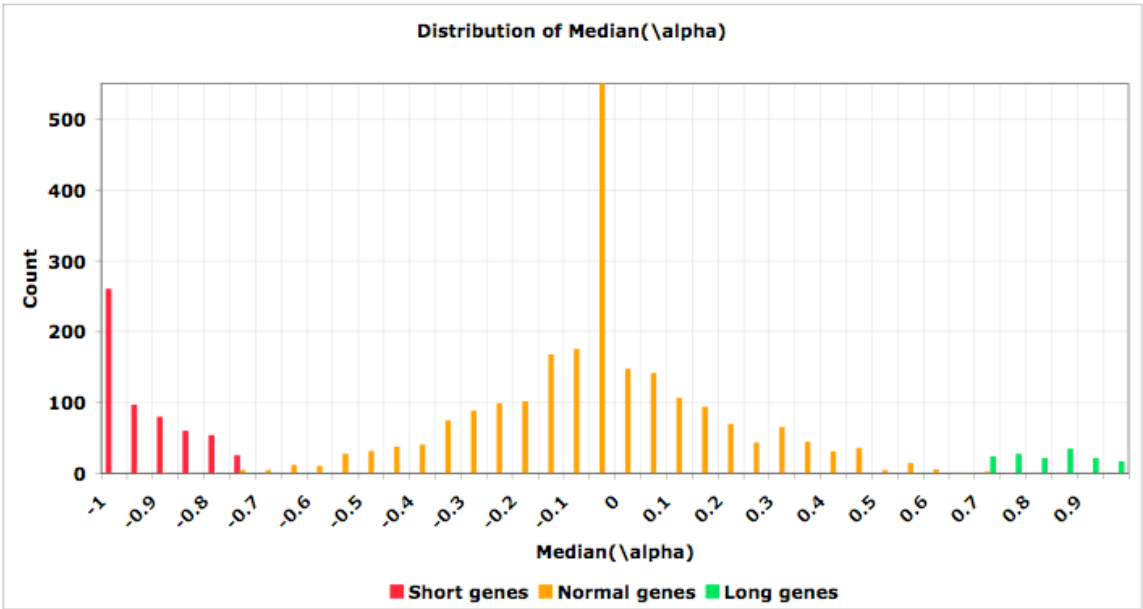
Identification of long, short, and matching genes based on average and median alignment scores (α).

Supplementary Figure 8: Distribution of mean alignment scores (α) in short, matching, and long genes



Distribution of mean alignment scores (α) in short, matching, and long genes. Data plotted is from gene calls made by Prodigal for five genomes with 68.3%, 56.4%, 58.65%, 42.83%, and 65.1% GC content, respectively.

Supplementary Figure 9: Distribution of median alignment scores (α) in short, matching, and long genes



Distribution of median alignment scores (α) in short, matching, and long genes. Data plotted is from gene calls made by Prodigal for five genomes with 68.3%, 56.4%, 58.65%, 42.83%, and 65.1% GC content, respectively.

Supplementary Table 1: Alignment scores (α) for short, matching, and long genes

		A = Mean(α)	B = Median(α)	A-B
Short	Mean	-0.8679038	-0.9094527	0.0457019
	STDV	0.11813249	0.08122121	0.06656605
Long	Mean	0.76133435	0.84391287	0.08340594
	STDV	0.11125017	0.07862425	0.07197461
Matching	Mean	-0.0372008	-0.0171879	0.06741398
	STDV	0.18803534	0.20608396	0.06731796

The mean and median values of α were calculated based on genes from five genomes that had been manually curated and identified as long, short, or matching genes.

Supplementary Data 1: Statistics of all public contigs processed by GenePRIMP

Contigs from ORNL	Genome GC%	CDSs	Short genes	Extended short genes	Long genes	Trimmed long genes	Unique genes	Dubious genes	Broken genes	Interupted genes	Missed genes	Added genes	Deleted genes	Pseudo genes	% corrections	Total number of anomalies	Date
Arthrobacter chlorophenolicus A6, Contig 2893	65.9	151	12	4	2	3	2	0	0	0	17	8	4	4	7.538	36	10/22/08
Methylobacterium nodulans ORS 2060, Contig 403	68.4	47	2	1	1	1	0	0	0	0	3	2	0	1	15.71	6	9/16/08
Methylobacterium chloromethanicum CM4, Contig 430	0	5355	213	0	176	0	104	2	24	19	649	0	0	0	0	1676	8/1/08
Methylobacterium chloromethanicum CM4, Contig 429	0	346	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8/1/08
Methylobacterium chloromethanicum CM4, Contig 428	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8/1/08
Methylobacterium nodulans ORS 2060, Contig 454	68.4	7643	326	222	162	110	24	2	56	0	329	172	116	301	15.71	899	8/13/08
Methylobacterium nodulans ORS 2060, Contig 400	68.4	8	0	1	0	0	0	0	0	0	1	2	2	1	15.71	1	8/13/08
Methylobacterium nodulans ORS 2060, Contig 401	68.4	13	0	0	0	0	0	0	0	0	0	1	1	0	15.71	0	8/13/08
Methylobacterium nodulans ORS 2060, Contig 402	68.4	21	1	0	0	0	0	0	2	0	1	2	0	2	15.71	4	8/13/08
Methylobacterium nodulans ORS 2060, Contig 404	68.4	59	7	6	0	1	0	0	0	0	11	3	2	7	15.71	18	8/13/08
Methylobacterium nodulans ORS 2060, Contig 406	68.4	456	52	34	21	16	3	0	8	6	107	45	19	72	15.71	217	8/13/08
Methylobacterium nodulans ORS 2060, Contig 407	68.4	487	54	44	24	13	1	1	24	13	119	45	28	101	15.71	248	8/13/08
Halorubrum lacusprofundi ATCC 49239, Contig 286	66.5	2749	75	31	127	54	34	0	9	6	29	8	9	33	8.55	274	8/15/08

Halorubrum lacusprofundi ATCC 49239, Contig 285	66.5	507	39	20	17	9	27	0	0	3	42	20	6	45	8.55	127	8/15/08
Halorubrum lacusprofundi ATCC 49239, Contig 284	66.5	393	39	0	11	20	27	0	2	0	37	20	11	26	8.55	116	8/15/08
Cyanothece sp. PCC 7424, Contig 955	0	5504	271	0	188	0	192	6	33	17	189	0	0	0	0	1202	9/9/08
Cyanothece sp. PCC 7424, Contig 954	0	303	6	0	1	0	18	0	0	2	30	0	0	0	0	63	9/9/08
Cyanothece sp. PCC 7424, Contig 953	0	216	32	0	5	0	18	1	0	2	18	0	0	0	0	93	9/9/08
Cyanothece sp. PCC 7424, Contig 952	0	28	4	0	3	0	2	0	0	0	4	0	0	0	0	13	9/9/08
Cyanothece sp. PCC 7424, Contig 951	0	21	1	0	1	0	4	0	0	0	0	0	0	0	0	7	9/9/08
Cyanothece sp. PCC 7424, Contig 950	0	21	1	0	1	0	1	0	0	0	0	0	0	0	0	7	9/9/08
Cyanothece sp. PCC 7424, Contig 949	0	11	0	0	2	0	1	0	0	0	0	0	0	0	0	3	9/9/08
Thauera sp. MZ1T, Contig 150	68.28	78	5	4	3	2	0	0	0	0	7	5	5	4	10.94	15	9/16/08
Dictyoglomus turgidum DSM 6724, Contig 33	33.96	1828	121	43	35	34	12	2	14	6	47	28	40	70	11.76	231	9/19/08
Desulfitobacterium hafniense DCB-2	47.54	2673	96	45	57	16	47	2	6	0	156	72	43	55	8.64	364	9/22/08
Cyanothece sp. PCC 8801, Contig 848	39.7	4448	156	135	74	75	84	2	38	22	197	105	103	186	14.22	559	9/22/08
Cyanothece sp. PCC 8801, Contig 847	39.7	46	2	1	0	0	1	0	0	2	4	4	2	4	14.22	8	9/22/08
Cyanothece sp. PCC 8801, Contig 846	39.7	63	8	6	0	1	3	0	0	0	13	7	9	9	14.22	24	9/22/08
Cyanothece sp. PCC 8801, Contig 845	39.7	13	0	0	0	0	1	0	0	0	2	1	1	1	14.22	3	9/22/08
Desulfitobacterium hafniense DCB-2, Part 2	47.54	2296	68	45	40	55	55	2	12	8	134	65	67	70	13.15	313	9/25/08
Methanosphaerula palustris E1-9c	55.35	2821	189	94	61	39	121	3	22	15	126	93	113	138	16.91	525	9/29/08
Desulfatibacillum alkenivorans AK-01	54.5	5298	168	73	96	48	107	1	23	10	88	31	35	40	4.285	556	9/29/08

Chloroflexus aggregans	56.4	4009	121	79	51	62	105	6	47	10	91	59	218	110	13.17	529	9/29/08
Shewanella baltica OS223, Contig 72	46.3	4389	94	77	34	35	61	2	25	15	266	41	89	57	7.123	558	9/30/08
Shewanella baltica OS223, Contig 71	46.3	64	4	4	0	0	3	0	0	0	4	1	0	2	7.123	11	9/30/08
Shewanella baltica OS223, Contig 70	46.3	50	0	0	0	0	0	0	0	0	0	0	0	0	7.123	0	9/30/08
Shewanella baltica OS223, Contig 69	46.3	88	4	2	1	1	4	0	0	0	11	6	5	7	7.123	22	9/30/08
Ammonifex degensii KC4, Contig 110	59.45	29	0	1	2	2	1	0	0	0	2	0	0	1	15.99	7	8/18/09
Geobacillus sp. Y412MC10, Contig 107	51.24	6332	151	110	133	86	115	0	38	31	144	71	58	103	6.76	693	8/21/09
Halothiobacillus neapolitanus c2, Contig 14	54.71	2433	66	42	81	33	65	2	21	0	73	47	64	55	9.9	366	8/21/09
Sulfolobus solfataricus 98/2, Contig 83	35.83	3032	128	171	45	32	50	2	111	131	317	124	237	224	26	906	8/21/09
Arthrobacter chlorophenolicus A6, Contig 2890	65.9	556	23	3	6	4	33	1	0	0	7	2	8	1	7.538	79	10/22/08
Arthrobacter chlorophenolicus A6, Contig 2883	65.9	3976	126	109	98	61	46	3	20	20	97	23	62	57	7.538	443	10/22/08
Cyanothece PCC 7425, Contig 116	50.6	29	0	1	0	0	0	0	0	0	0	0	1	0	8.659	0	10/22/08
Cyanothece PCC 7425, Contig 117	50.6	207	26	14	5	5	20	0	8	2	22	5	5	25	8.659	87	10/22/08
Cyanothece PCC 7425, Contig 118	50.6	177	17	5	8	5	18	0	2	0	9	2	10	8	8.659	61	10/22/08
Cyanothece PCC 7425, Contig 120	50.6	5119	178	63	113	82	155	3	46	16	137	56	99	93	8.659	743	10/22/08
Anaeromyxobacter dehalogenans strain 2CP-1, Contig 193	74.72	4523	65	39	24	19	3	0	23	6	212	21	28	39	3.228	394	10/23/08
Methanocaldococcus vulcanius M7, Contig 80	31.49	2	0	0	0	0	0	0	0	0	0	0	0	0	8.63	0	8/21/09
Desulfovibrio desulfuricans 27774	58.07	2410	51	19	73	55	46	3	10	2	30	9	33	26	5.892	281	11/5/08

Thioalkalivibrio sp. HL-EbGR7	65.1	3367	154	94	48	38	41	1	37	24	131	64	108	106	12.18	463	11/11/08
Clostridium cellulolyticum H10	37.4	3546	101	62	44	44	96	2	22	14	124	75	128	98	11.48	513	11/17/08
Diaphorobacter sp. TPSY	66.83	3528	127	118	40	49	14	1	6	10	252	61	42	74	9.751	481	11/20/08
Geobacillus sp. WCH70, Contig 196	43.83	3468	152	119	42	78	57	1	43	47	328	181	203	308	25.74	777	11/20/08
Geobacillus sp. WCH70, Contig 193	43.83	33	2	2	0	1	0	0	0	0	4	3	1	3	25.74	8	11/20/08
Geobacillus sp. WCH70, Contig 192	43.83	11	1	1	0	0	1	0	0	0	0	1	3	0	25.74	4	11/20/08
Anaerocellum thermophilum, Contig 268	35.17	2829	90	47	31	23	52	4	42	23	190	57	110	98	11.82	520	12/1/08
Anaerocellum thermophilum, Contig 263	35.17	8	0	1	1	0	1	0	0	0	0	0	0	0	11.82	2	12/1/08
Anaerocellum thermophilum, Contig 259	35.17	4	1	0	0	0	0	0	0	0	0	0	0	0	11.82	1	12/1/08
Geobacter sp. FRC-32	53.47	3837	85	41	52	44	56	1	32	2	151	34	31	41	4.98	455	12/8/08
Ralstonia pickettii 12D, Contig 80	63.56	3467	94	64	54	48	42	2	21	11	387	24	30	49	6.59	652	12/9/08
Ralstonia pickettii 12D, Contig 79	63.56	1204	42	23	22	15	16	1	6	0	77	12	11	24	6.59	185	12/9/08
Ralstonia pickettii 12D, Contig 78	63.56	431	28	6	3	6	32	0	2	0	20	2	1	5	6.59	93	12/9/08
Ralstonia pickettii 12D, Contig 77	63.56	289	15	9	6	4	14	1	2	4	36	5	7	11	6.59	82	12/9/08
Ralstonia pickettii 12D, Contig 76	63.56	69	7	2	0	0	4	0	0	0	7	0	1	1	6.59	18	12/9/08
Sulfolobus islandicus U.3.28, Contig 145	34.64	3175	113	80	41	56	68	1	93	43	266	84	167	170	17.54	766	1/13/09
HRB1 sp., Contig 19	49.03	3193	59	33	39	27	31	0	4	8	96	31	60	41	6.01	329	1/26/09
Rhizobium leguminosarum bv. trifolii WSM1325, Contig 1172	61.09	316	7	6	2	4	7	0	4	2	20	11	6	12	10.78	89	2/6/09
Rhizobium leguminosarum bv. trifolii WSM1325, Contig 1165	61.09	529	39	28	10	10	33	0	8	6	73	54	21	64	10.78	180	2/6/09

Rhizobium leguminosarum bv. trifolii WSM1325, Contig 1171	61.09	312	8	5	4	2	6	0	2	2	27	9	4	9	10.78	67	2/6/09
Rhizobium leguminosarum bv. trifolii WSM1325, Contig 1170	61.09	691	35	25	9	9	17	0	6	7	73	29	14	37	10.78	262	2/6/09
Rhizobium leguminosarum bv. trifolii WSM1325, Contig 1169	61.09	786	25	25	15	13	7	1	12	12	70	24	15	31	10.78	229	2/6/09
Rhizobium leguminosarum bv. trifolii WSM1325, Contig 1168	61.09	4642	77	73	84	59	78	0	30	24	347	66	45	74	10.78	730	2/6/09
Variovorax paradoxus S110, Contig 49	67.63	1052	44	25	23	20	16	1	6	0	47	37	9	42	7.19	146	2/19/09
Exiguobacterium sp. AT1b, Contig 100	48.46	3078	66	22	28	27	46	4	7	10	27	11	46	24	4.22	268	2/19/09
Variovorax paradoxus S110, Contig 54	67.63	5264	167	105	66	67	47	3	25	6	124	61	22	66	7.19	487	2/19/09
Thauera sp. MZ1T, Contig 151	68.4	3992	116	104	97	73	28	1	45	40	156	58	71	109	10.78	500	2/28/09
Thermotogales bacterium TBF 19.5.1, Contig 147	41.55	2191	57	35	32	35	24	0	38	12	45	25	42	51	8.35	286	3/10/09
Pectobacterium carotovorum subsp. wasabiae, Contig 316	50.48	4526	113	80	97	64	76	1	29	19	353	157	62	178	11.95	733	7/7/09
Dickeya dadantii Ech1591, Contig 74	54.52	4229	108	80	83	63	67	0	21	17	270	92	43	104	9.03	607	3/11/09
Fibrobacter succinogenes S85 ATCC 19169, Contig 123	48.05	3133	120	40	79	24	60	1	13	18	31	26	30	34	4.9	358	8/14/09
Ammonifex degensii KC4, Contig 140	59.45	2191	117	83	27	34	29	5	50	58	77	51	82	101	15.99	494	8/18/09
Pectobacterium carotovorum ssp. carotovorum PC1, Contig 58	51.93	4292	70	60	84	49	59	1	22	12	218	45	37	49	5.59	506	3/31/09
Paenibacillus sp. JDR-2, Contig 301	50.28	6307	163	70	91	63	83	0	47	43	103	44	60	70	4.87	586	3/31/09
Desulfovibrio salexigens DSM 2638, Contig 92	47.09	3857	85	55	47	29	35	1	41	27	37	17	39	25	4.28	326	4/9/09
Geobacter sp. M21, Contig 449	60.47	4109	84	59	46	41	28	0	39	18	196	72	27	72	6.6	442	4/14/09

Escherichia coli BL21(DE3), Contig 2859	50.84	4308	125	135	6	13	4	1	58	64	1026	137	101	99	11.26	1343	4/21/09
Methanocaldococcus fervens AG86 , Contig 452	32.21	35	1	1	1	1	1	0	0	0	0	0	0	0	11.3	4	7/17/09
Methanocaldococcus fervens AG86 , Contig 453	32.21	1622	55	46	32	28	15	0	40	22	58	21	51	39	11.3	261	7/17/09
Methanococcus voltae A3, Contig 1958	28.59	1740	37	29	135	47	21	0	15	12	22	5	20	3	5.98	252	7/23/09
Bacillus selenitireducens MLS-10, ATCC 700615, Contig 105	48.67	3377	63	78	48	36	54	1	84	80	138	37	88	74	9.27	473	7/28/09
Zymomonas mobilis pomaceae lectotype ATCC 29192, Contig 34	44.09	42	1	2	0	0	1	0	4	4	6	3	8	3	6.25	16	8/4/09
Zymomonas mobilis pomaceae lectotype ATCC 29192, Contig 35	44.09	1701	37	27	31	15	10	0	12	14	57	18	18	15	6.25	174	8/4/09
Lutiella nitroferrum 2002, Contig 81	64.57	3905	116	54	53	44	45	4	0	7	0	13	25	25	4.12	265	8/6/09
Methanocaldococcus vulcanius M7, Contig 84	31.49	13	0	1	0	0	2	0	0	0	1	1	0	1	8.63	4	8/21/09
Methanocaldococcus vulcanius M7, Contig 94	31.49	1769	47	31	44	31	36	1	16	8	57	32	35	22	8.63	260	8/24/09
Geobacillus sp. Y412MC61, Contig 130	52.42	40	2	0	1	0	1	0	0	0	7	1	1	1	14.57	11	8/27/09
Geobacillus sp. Y412MC61, Contig 166	52.42	3577	90	82	65	59	47	0	64	55	363	107	115	163	14.57	789	8/28/09
Victivallis vadensis ATCC BAA-548, permanent draft	59	4129	141	53	88	26	54	3	22	18	53	36	32	59	4.99	445	9/2/09
Clostridium thermocellum DSM 4150, Permanent draft	39	3345	95	101	45	30	72	1	100	20	241	116	191	182	18.55	665	9/2/09
Nostoc azollae 0708, Contig 953	38.45	13	1	2	0	0	3	0	0	0	5	1	0	3	82	9	9/20/09
Nostoc azollae 0708, Contig 1927	38.45	159	21	41	0	1	16	1	60	2	74	21	66	63	82	172	9/20/09
Nostoc azollae 0708, Contig 2627	38.45	6726	608	987	129	114	468	9	1877	388	1951	622	2118	1620	82	5252	9/23/09

Desulfonatronospira thiodismutans ASO3-1	0	3794	114	0	59	0	190	0	47	30	266	0	0	0	0	803	1/27/10
Thermotoga naphthophila RKU-10, Contig 4	0	1827	41	0	17	0	5	0	44	36	96	0	0	0	0	346	9/30/09
Thermoanaerobacter italicus Ab9 DSM 9252, Contig 360	34.14	2386	44	56	18	27	31	0	63	20	210	96	70	137	16.18	429	11/10/09
Dickeya dadantii Ech586, Contig 42	53.64	4192	56	68	62	31	77	1	18	4	279	72	44	72	6.85	543	10/22/09
Zymomonas mobilis subsp. mobilis ATCC 10988, Contig 11	46.22	2	0	0	0	0	0	0	0	0	0	0	0	0	15.64	0	10/26/09
Zymomonas mobilis subsp. mobilis ATCC 10988, Contig 12	46.22	3	0	0	0	0	0	0	0	0	0	0	0	0	15.64	0	10/26/09
Zymomonas mobilis subsp. mobilis ATCC 10988, Contig 13	46.22	31	2	1	0	2	0	0	2	0	12	9	2	10	15.64	16	10/26/09
Zymomonas mobilis subsp. mobilis ATCC 10988, Contig 14	46.22	32	2	1	0	2	2	0	2	0	6	8	8	8	15.64	12	10/26/09
Zymomonas mobilis subsp. mobilis ATCC 10988, Contig 15	46.22	24	0	0	1	0	1	0	0	0	7	3	0	3	15.64	9	10/26/09
Zymomonas mobilis subsp. mobilis ATCC 10988, Contig 16	46.22	28	1	1	0	0	0	0	0	0	10	6	2	5	15.64	11	10/26/09
Allochromatium vinosum DSM 180, Contig 249	64.37	44	0	0	1	2	2	0	0	0	6	5	0	1	9.05	11	10/27/09
Dehalococcoides sp. VS, Contig 05	0	1459	25	0	17	0	19	0	2	0	86	0	0	0	0	175	10/27/09
Allochromatium vinosum DSM 180, Contig 250	64.37	126	6	5	4	0	9	0	2	0	14	7	4	8	9.05	38	10/27/09
Zymomonas mobilis subsp. mobilis ATCC 10988, Contig 17	46.22	1747	33	43	29	26	12	0	11	6	289	54	58	40	15.64	408	10/27/09
Allochromatium vinosum DSM 180, Contig 251	64.37	3091	66	59	58	41	37	1	13	4	116	64	27	72	9.05	334	10/27/09

Natrialba magadii ATCC 43099, Contig 64	61.42	3617	38	32	143	51	223	0	16	8	154	45	52	50	8.1	613	1/6/10
Natrialba magadii ATCC 43099, Contig 63	61.42	352	9	7	14	9	21	0	1	2	16	6	4	13	8.1	70	1/5/10
Natrialba magadii ATCC 43099, Contig 62	61.42	239	10	18	4	3	15	0	2	2	28	18	9	29	8.1	60	1/5/10
Methylotenera sp. 301, Contig 1822	42.64	2795	47	51	48	36	43	0	11	4	50	20	14	35	5.58	226	10/28/09
Ferroglobus placidus DSM 10642, Contig 489	44.14	2589	66	55	21	17	95	0	36	5	71	51	68	87	10.74	415	10/28/09
Natrialba magadii ATCC 43099, Contig 61	61.42	94	1	0	2	1	3	0	0	0	2	0	0	0	8.1	10	1/5/10
Starkeya novella DSMZ, Contig 92	67.88	4480	96	88	69	47	89	0	26	2	240	55	24	78	5.89	565	11/5/09
Sideroxydans lithotrophicus ES-1, Contig 246	57.54	2972	76	66	28	21	0	0	14	8	82	39	14	16	5.25	237	11/24/09
Thioalkalivibrio sp. K90mix, Contig 474	65.87	283	5	2	4	4	26	0	2	0	19	4	2	5	5.23	63	12/21/09
Thioalkalivibrio sp. K90mix, Contig 352	65.87	2602	36	39	40	29	26	0	14	19	75	18	20	28	5.23	237	12/21/09
Klebsiella variicola At-22, Contig 38	57.58	5038	62	61	41	42	20	0	5	0	720	73	10	41	4.5	905	12/23/09
Methanocaldococcus sp. FS406-22, Contig 94	32.04	13	2	1	0	0	0	0	0	0	1	2	0	0	9.64	4	1/14/10
Methanocaldococcus sp. FS406-22, Contig 92	32.04	1854	30	34	25	24	25	0	33	14	51	40	47	32	9.64	226	1/14/10
Zymomonas mobilis pomaceae lectotype ATCC 29192, Contig 56	43	42	2	1	0	1	3	0	4	0	5	5	10	4	56.98	18	1/14/10
Zymomonas mobilis pomaceae lectotype ATCC 29192, Contig 54	43	44	2	4	0	0	6	0	5	0	15	6	9	9	56.98	28	1/14/10
Desulfurivibrio alkaliphilus AHT2, Contig 27	60.29	2690	66	66	44	15	26	0	24	14	294	35	49	53	8.1	502	1/14/10
Dehalococcoides sp. GT, Contig 14	47.31	1428	20	27	13	11	18	0	10	6	97	18	12	18	6.02	184	1/19/10

Alicyclobacillus acidocaldarius LAA1	0	2981	104	0	37	0	106	0	51	20	272	0	0	0	0	652	1/23/10
Burkholderia sp. CCGE1002, Contig 96	0	487	35	0	4	0	27	0	37	13	100	0	0	0	0	214	2/2/10
Burkholderia sp. CCGE1002, Contig 97	0	1180	43	0	15	0	33	0	30	4	133	0	0	0	0	274	2/2/10
Burkholderia sp. CCGE1002, Contig 99	0	3162	56	0	51	0	27	0	16	10	287	0	0	0	0	465	2/3/10
Burkholderia sp. CCGE1002, Contig 98	0	2333	61	0	45	0	25	0	31	12	219	0	0	0	0	404	2/3/10
Caulobacter segnis ATCC 21756, Contig 532	0	4293	114	0	39	0	55	0	89	79	414	0	0	0	0	768	2/5/10
Contigs from GEBA	**Genome GC%**	**CDSs**	**Short genes**	**Extended short genes**	**Long genes**	**Trimmed long genes**	**Unique genes**	**Dubious genes**	**Broken genes**	**Interupted genes**	**Missed genes**	**Added genes**	**Deleted genes**	**Pseudo genes**	**% corrections**	**Total number of anomalies**	**Date**
Sanguibacter keddieii DSM 10542	71.9	3746	222	162	182	56	57	8	19	2	68	30	41	25	8.356	556	9/19/08
Slackia heliotrinireducens DSM 20476	60.2	2819	119	55	61	23	65	7	22	2	45	15	36	33	5.747	320	9/23/08
Cryptobacterium curtum DSM 15641	50.9	1367	38	15	77	36	29	0	4	2	27	5	8	7	5.194	175	9/25/08
Saccharomonospora viridis P101, DSM 43017	67.3	3959	292	274	74	89	76	10	19	10	148	43	96	78	14.65	676	9/29/08
Kytococcus sedentarius DSM 20547	71.6	2662	184	76	81	39	46	12	13	30	108	42	65	84	11.5	478	10/2/08
Sulfurimonas autotrophica DSM 16294, Contig 12	35.24	2175	46	21	14	16	13	0	0	2	0	1	11	8	2.62	130	8/21/09
Desulfomicrobium baculatum DSM 4028	58.65	3451	80	64	45	22	56	2	16	22	101	63	20	58	6.58	365	12/10/08
Leptotrichia buccalis DSM 1135	29.65	2335	59	45	42	30	40	1	62	48	68	58	84	91	13.19	323	12/10/08
Capnocytophaga ochracea DSM 7271	39.59	2216	58	33	23	9	61	0	9	10	34	14	35	20	5.01	219	12/16/08

Acidimicrobium Ferrooxidans DSM 10331	68.3	2070	105	43	29	25	34	3	23	17	63	44	76	74	12.31	335	12/17/08
Actinosynnema mirum DSM 43827	73.71	7107	338	259	135	65	121	3	64	25	223	81	324	179	9.43	943	12/17/08
Beutenbergia cavernae DSM 12333	73.1	4222	118	31	93	52	20	2	20	6	52	14	13	30	3.31	335	12/17/08
Catenulispora acidiphila DSM 44928	69.8	9075	354	178	241	83	162	9	56	40	275	100	117	144	6.85	1232	12/17/08
Halomicrobium mukohataei DSM 12286, Contig 62	65.6	3205	85	33	76	43	55	2	6	8	79	41	21	46	6.75	322	12/18/08
Halomicrobium mukohataei DSM 12286, Contig 61	65.6	190	9	7	8	2	8	2	2	0	12	12	6	18	6.75	42	12/18/08
Dyadobacter fermentans DSM 18053	51.5	5851	230	112	142	80	144	3	71	51	161	53	100	88	7.4	827	12/18/08
Halorhabdus utahensis DSM 12940	62.9	3047	90	24	88	62	70	2	15	15	72	19	39	29	5.68	364	12/19/08
Meiothermus ruber DSM 1279	63.4	3083	114	52	40	27	55	0	22	23	58	22	51	40	6.23	353	12/19/08
Pedobacter heparinus DSM 2366	42	4314	122	73	72	49	74	3	34	43	79	30	59	41	5.84	456	12/19/08
Anaerococcus prevotii DSM 20548, Contig 723	36.1	1735	60	29	16	8	18	0	19	12	47	20	20	38	7.66	177	12/19/08
Anaerococcus prevotii DSM 20548, Contig 698	36.1	106	5	5	0	1	5	0	2	0	15	9	2	9	7.66	30	12/19/08
Sphaerobacter thermophilus DSM 20745	68.1	2466	91	30	48	26	42	2	6	8	36	10	15	24	4.26	266	12/22/08
Thermobispora bispora DSM 43833	72.4	3604	169	96	55	38	55	2	28	26	107	42	47	48	7.52	477	12/22/08
Pirellula staleyi DSM 6068	57.5	4819	106	38	163	102	172	5	38	12	65	27	71	54	6.06	642	12/22/08
Eggerthella lenta DSM 02243	64.2	3113	86	40	100	60	54	3	21	18	61	32	22	52	6.62	357	12/22/08
Planctomyces limnophilus DSM 3776, Contig 476	54	4292	87	27	149	116	235	4	15	8	51	21	68	47	6.5	659	12/22/08
Planctomyces limnophilus DSM 3776, Contig 422	54	61	4	1	0	0	9	0	0	0	1	1	2	0	6.5	16	12/22/08
Streptosporangium roseum DSM 43021, Contig 236	70.87	9281	448	402	191	139	205	5	59	85	431	211	101	435	13.91	1541	12/22/08
Streptosporangium roseum DSM 43021, Contig 206	70.87	33	5	1	0	0	1	0	3	2	2	2	3	2	13.91	11	12/22/08

Atopobium parvulum DSM 20469, Contig 22	45.69	1367	26	13	65	20	21	0	1	2	26	9	7	16	4.75	154	1/13/09
Jonesia denitrificans DSM 20603, Contig 118	58.42	2596	90	48	68	72	86	1	16	10	44	19	59	47	9.44	403	1/13/09
Rhodothermus marinus DSM 4252, Contig 87	64.54	112	12	8	4	4	6	1	3	0	11	13	10	10	7.01	45	1/14/09
Kangiella koreensis DSM 16069, Contig 47	43.69	2655	64	37	49	36	35	1	2	4	35	13	23	16	4.71	225	1/14/09
Veillonella parvula DSM 2008, Contig 46	38.63	1865	38	20	25	15	21	2	18	16	17	17	15	16	4.45	154	1/14/09
Rhodothermus marinus DSM 4252, Contig 88	64.54	2811	74	34	52	31	25	0	26	22	43	25	27	43	7.01	302	1/14/09
Xylanimonas cellulosilytica DSM 15894, Contig 38	72.5	104	6	1	1	0	2	0	0	0	4	2	0	0	5.66	13	1/14/09
Streptobacillus moniliformis DSM 12112, Contig 187	26.31	8	0	0	0	0	0	0	0	0	0	0	0	0	14.62	0	1/14/09
Thermobaculum terrenum ATCC BAA-798, Contig 144	63.76	1004	50	17	17	16	27	2	12	14	20	14	19	22	6.65	149	1/14/09
Streptobacillus moniliformis DSM 12112, Contig 186	26.31	1538	73	45	19	12	26	0	47	23	58	35	65	69	14.62	265	1/14/09
Thermobaculum terrenum ATCC BAA-798, Contig 145	48.07	1895	53	32	39	21	38	0	8	4	33	10	25	17	6.65	207	1/14/09
Xylanimonas cellulosilytica DSM 15894, Contig 570	72.5	3375	114	65	67	41	37	1	22	10	79	24	23	41	5.66	341	1/14/09
Chitinophaga pinensis DSM 2588, Contig 402	45.23	7332	219	110	147	96	223	5	75	37	141	82	114	110	6.98	939	1/27/09
Desulfotomaculum acetoxidans DSM 771, Contig 96	41.55	4368	194	137	79	88	210	4	110	78	235	196	178	304	20.67	1002	1/28/09
Sulfurospirillum deleyianum DSM 6946, Contig 103	39	2294	49	27	13	14	21	0	6	8	26	14	16	27	4.27	170	8/17/09
Thermomonospora curvata DSM 43183, Contig 146	71.64	5009	211	107	74	60	100	4	60	39	131	58	68	99	7.83	678	2/3/09
Thermanaerovibrio acidaminovorans DSM 6589, Contig 128	63.79	1764	69	45	24	9	14	0	21	17	29	23	22	27	7.14	211	2/19/09
Stackebrandtia nassauensis DSM 44728, Contig 197	68.13	6456	260	121	108	94	95	2	21	13	154	75	51	108	6.95	809	2/20/09

Haloterrigena turkmenica DSM 5511, Contig 83	65.8	195	18	14	4	6	17	1	3	0	46	38	15	43	10.7	92	7/9/09
Haloterrigena turkmenica DSM 5511, Contig 81	65.8	159	5	1	7	0	2	0	0	2	2	1	0	0	10.7	19	7/9/09
Haloterrigena turkmenica DSM 5511, Contig 82	65.8	96	6	2	0	0	8	0	0	0	4	2	0	1	10.7	18	7/9/09
Haloterrigena turkmenica DSM 5511, Contig 80	65.8	21	0	0	0	0	1	0	0	0	1	1	0	1	10.7	2	7/9/09
Methanohalophilus mahii DSM 5219, Contig 30	42.62	2035	64	31	35	32	34	1	48	22	72	30	29	45	8.21	291	3/11/09
Nakamurella multipartita DSM 44233, Contig 351	70.92	5486	277	157	177	134	57	3	117	136	207	84	141	174	12.58	996	3/11/09
Archaeoglobus profundus DSM 5631, Contig 188	39.84	4	0	0	0	0	0	0	0	0	0	0	0	0	6.93	0	3/17/09
Spirosoma linguale DSM 74, Contig 241	50.19	41	3	2	1	1	4	0	0	2	2	0	0	1	8.47	14	3/17/09
Spirosoma linguale DSM 74, Contig 233	50.19	10	0	0	0	0	1	0	0	0	0	0	0	0	8.47	1	3/17/09
Spirosoma linguale DSM 74, Contig 224	50.19	9	2	0	0	0	1	0	0	0	0	0	1	0	8.47	3	3/17/09
Spirosoma linguale DSM 74, Contig 225	50.19	11	0	0	0	0	0	0	0	0	0	0	0	0	8.47	0	3/17/09
Spirosoma linguale DSM 74, Contig 226	50.19	14	0	0	0	1	0	0	0	0	0	0	0	0	8.47	2	3/17/09
Spirosoma linguale DSM 74, Contig 227	50.19	11	2	1	0	0	1	0	0	0	1	1	0	0	8.47	4	3/17/09
Sebaldella termitidis ATCC 33386, Contig 68	33.45	16	0	0	0	0	0	0	0	0	0	0	0	0	4.59	0	3/17/09
Spirosoma linguale DSM 74, Contig 247	50.19	185	8	4	5	4	10	1	4	2	9	8	8	8	8.47	39	3/17/09
Sebaldella termitidis ATCC 33386, Contig 69	33.45	53	4	1	1	1	1	0	0	0	3	3	1	4	4.59	12	3/17/09
Spirosoma linguale DSM 74, Contig 234	50.19	152	3	3	2	3	5	0	6	4	5	2	7	5	8.47	27	3/17/09
Archaeoglobus profundus DSM 5631, Contig 189	39.84	1872	77	29	24	26	46	1	23	20	28	11	29	35	6.93	300	3/17/09

Sebaldella termitidis ATCC 33386, Contig 70	33.45	4139	88	43	39	17	76	4	42	22	79	34	33	56	4.59	387	3/18/09
Spirosoma linguale DSM 74, Contig 279	50.19	6695	199	120	202	133	195	5	75	70	133	61	112	118	8.47	974	3/18/09
Conexibacter woesei DSM 14684, Contig 121	72.73	5941	302	160	153	30	45	0	12	19	78	21	12	36	4.36	669	7/7/09
Denitrovibrio acetiphilus DSM 12809, Contig 92	42.54	3023	98	55	18	18	17	0	19	37	58	49	49	59	7.6	302	7/2/09
Haloterrigena turkmenica DSM 5511, Contig 84	65.8	350	14	10	8	3	28	0	6	2	39	36	18	39	10.7	96	7/10/09
Haliangium ochraceum DSM 14365, Contig 199	69.48	6888	304	218	228	87	84	0	39	58	227	66	54	178	8.75	1011	4/9/09
Desulfohalobium retbaense DSM 5692, Contig 73	57.5	51	4	5	1	0	3	0	2	0	6	3	2	6	5.87	18	4/20/09
Gordonia bronchialis DSM 43247, Contig 180	67.07	78	2	0	1	1	2	0	0	0	5	1	0	0	19.03	10	4/20/09
Desulfohalobium retbaense DSM 5692, Contig 76	57.6	2506	43	33	46	32	44	1	7	12	56	22	28	19	5.87	295	4/21/09
Gordonia bronchialis DSM 43247, Contig 194	67.07	4955	359	290	186	146	59	3	144	142	231	98	179	243	19.03	1096	4/21/09
Alicyclobacillus acidocaldarius DSM 446, Contig 108	62.33	5	0	0	0	0	0	0	0	0	0	0	0	0	9.56	0	4/22/09
Alicyclobacillus acidocaldarius DSM 446, Contig 109	62.33	99	10	5	1	3	2	1	3	0	13	3	3	7	9.56	34	4/22/09
Alicyclobacillus acidocaldarius DSM 446, Contig 110	62.33	98	7	0	1	3	2	0	0	0	10	6	2	3	9.56	24	4/22/09
Alicyclobacillus acidocaldarius DSM 446, Contig 111	62.33	2956	78	56	49	38	19	0	30	33	139	45	46	65	9.56	407	4/22/09
Kribbella flavida DSM 17836	70.57	7079	406	290	138	56	56	2	45	45	215	64	57	139	8.56	955	5/5/09
Haloterrigena turkmenica DSM 5511, Contig 85	65.8	625	39	28	27	18	21	0	10	11	54	37	18	51	10.7	158	7/10/09
Haloterrigena turkmenica DSM 5511, Contig 86	65.8	3779	113	44	107	49	63	4	10	17	79	20	21	42	10.7	406	7/10/09
Tsukamurella paurometabola DSM 20162, Contig 105	68.41	92	5	3	0	3	3	0	4	0	7	5	5	7	10.08	19	7/22/09
Thermosphaera aggregans DSM 11486, Contig 48	46.73	1425	47	26	21	25	17	1	18	12	26	12	26	23	7.85	184	7/22/09

Tsukamurella paurometabola DSM 20162, Contig 106	68.41	4254	223	164	100	62	39	4	32	50	122	46	57	86	10.08	595	7/22/09
Aminobacterium colombiense DSM 12261, Contig 324	45.31	1921	54	31	34	28	45	0	12	14	39	25	31	38	7.96	252	7/23/09
Ignisphaera aggregans DSM 17230	35.69	2029	69	54	45	33	87	2	22	16	41	31	59	62	11.78	344	7/28/09
Sphaerobacter thermophilus DSM 20745, Contig 4369	68.12	1069	33	26	28	8	10	0	10	10	33	9	14	16	6.83	133	7/31/09
Ferrimonas balearica DSM 9799, Contig 117	60.22	3799	100	76	38	37	30	5	6	10	71	15	10	21	4.19	300	7/31/09
Geodermatophilus obscurus DSM 43160	73.98	5116	346	198	179	101	78	1	80	33	286	98	299	247	14.89	1065	8/5/09
Thermocrinis albus DSM 14484, Contig 4	46.93	1632	45	26	5	6	34	2	5	8	29	12	37	10	5.58	246	8/22/09
Spirochaeta smaragdinae DSM 11293	0	4415	130	0	81	0	83	1	152	26	108	0	0	0	0	689	9/14/09
Methanothermus fervidus DSM 2088, Contig 16	31.64	1318	27	30	15	6	9	0	22	16	16	8	23	18	6.45	135	11/10/09
Ilyobacter polytropus DSM 2926, Contig 58	34.53	125	9	4	1	4	1	0	4	0	13	12	10	19	8.66	29	10/9/09
Ilyobacter polytropus DSM 2926, Contig 59	34.53	912	26	18	13	4	30	0	14	16	26	15	27	23	8.66	118	10/10/09
Ilyobacter polytropus DSM 2926, Contig 60	34.53	1929	49	35	30	9	48	2	28	33	29	14	34	29	8.66	216	10/10/09
Acetohalobium arabaticum DSM 5501, Contig 136	36.63	2353	66	57	28	28	41	3	36	36	81	39	36	71	9.82	300	10/10/09
Acidaminococcus fermentans DSM 20731, Contig 169	55.84	2054	55	46	15	12	24	1	19	5	78	61	20	66	9.98	232	10/12/09
Olsenella uli DSM 7084	64.7	1773	32	33	74	23	1	0	5	2	76	30	16	47	8.4	213	11/20/09
Segniliparus rotundus DSM 44985, Contig 67	66.79	3071	102	93	60	30	119	0	26	4	138	38	35	68	8.6	522	11/9/09
Bacillus tusciae DSM 2912, Contig 79	59.11	3320	126	99	69	83	102	2	40	20	230	92	85	173	16.02	745	10/27/09
Brachyspira murdochii DSM 12563, Contig 316	27.75	2867	43	25	23	10	96	0	42	4	92	19	32	45	4.57	328	10/28/09
Cellulomonas flavigena DSM 20109, Contig 371	74.29	3721	84	56	84	48	78	0	18	0	256	17	8	52	4.86	561	10/28/09

Arcobacter nitrofigilis DSM 7299, Contig 168	28.36	3144	42	36	26	14	4	0	14	10	37	17	11	24	3.24	157	11/24/09
Spirochaeta smaragdinae DSM 11293, Contig 531	0	4331	103	0	77	0	0	0	60	52	99	0	0	0	0	440	11/29/09
Coraliomargarita akajimensis DSM 45221	53.6	3145	62	51	97	57	101	0	2	11	63	9	13	16	4.64	389	12/17/09
Nocardiopsis dassonvillei DSM 43111, Contig 412	0	703	14	0	16	0	21	0	0	0	72	0	0	0	0	128	12/22/09
Nocardiopsis dassonvillei DSM 43111, Contig 768	0	4859	129	0	113	0	121	0	20	10	589	0	0	0	0	1024	12/22/09
Meiothermus silvanus DSM 9946, Contig 2050	0	341	20	0	5	0	15	0	8	2	28	0	0	0	0	92	1/21/10
Meiothermus silvanus DSM 9946, Contig 1115	0	139	7	0	1	0	17	0	0	0	4	0	0	0	0	34	1/21/10
Meiothermus silvanus DSM 9946, Contig 2054	0	3229	95	0	53	0	80	0	39	35	125	0	0	0	0	476	1/22/10
Ktedonobacter racemifer DSM 44963	0	13083	551	0	418	0	1275	0	468	396	851	0	0	0	0	4339	2/1/10
Aminomonas paucivorans DSM 12260, Contig 78	0	2429	61	0	32	0	34	0	7	11	93	0	0	0	0	273	2/4/10
Contigs from GBP	**Genome GC%**	**CDSs**	**Short genes**	**Extended short genes**	**Long genes**	**Trimmed long genes**	**Unique genes**	**Dubious genes**	**Broken genes**	**Interupted genes**	**Missed genes**	**Added genes**	**Deleted genes**	**Pseudo genes**	**% corrections**	**Total number of anomalies**	**Date**
Mycobacterium sp. Spyr1, Chromosome (GeneMark)	0	5321	506	0	73	0	45	11	20	52	558	0	0	0	0	1326	10/28/08
Mycobacterium sp. Spyr1, Chromosome (Prodigal)	67.9	5302	311	308	67	76	23	2	20	55	416	89	98	153	15.4	943	10/29/08
Mycobacterium sp. Spyr1, Plasmid 1	67.9	234	20	17	6	8	10	UNAV	0	0	37	17	28	26	15.4	81	11/11/08
Mycobacterium sp. Spyr1, Plasmid 2	67.9	48	3	6	2	5	19	UNAV	0	0	7	2	24	3	15.4	32	11/11/08
Pichia stipitis CBS 6054, chromosome 1	UNAVAI	1276	64	UNAVA	16	UNAVA	38	UNAV	5	2	164	UNAV	UNAV	UNAV	UNAVA	289	1/12/09

Pichia stipitis CBS 6054, chromosome 2	UNAVAI	1046	69	UNAVA	15	UNAVA	19	UNAV	0	0	117	UNAV	UNAV	UNAV	UNAVA	220	1/12/09
Pichia stipitis CBS 6054, chromosome 3	UNAVAI	690	40	UNAVA	11	UNAVA	13	UNAV	0	2	97	UNAV	UNAV	UNAV	UNAVA	163	1/12/09
Pichia stipitis CBS 6054, chromosome 4	UNAVAI	693	43	UNAVA	5	UNAVA	16	UNAV	2	4	93	UNAV	UNAV	UNAV	UNAVA	163	1/12/09
Pichia stipitis CBS 6054, chromosome 5	UNAVAI	679	44	UNAVA	10	UNAVA	21	UNAV	2	6	84	UNAV	UNAV	UNAV	UNAVA	164	1/12/09
Pichia stipitis CBS 6054, chromosome 6	UNAVAI	685	41	UNAVA	4	UNAVA	16	UNAV	0	2	83	UNAV	UNAV	UNAV	UNAVA	146	1/12/09
Pichia stipitis CBS 6054, chromosome 7	UNAVAI	383	19	UNAVA	3	UNAVA	8	UNAV	0	0	6	UNAV	UNAV	UNAV	UNAVA	36	1/12/09
Pichia stipitis CBS 6054, chromosome 8	UNAVAI	364	30	UNAVA	2	UNAVA	9	UNAV	0	0	53	UNAV	UNAV	UNAV	UNAVA	94	1/12/09
Methanosphaerula palustris E1-9c	UNAVAI	3177	115	UNAVA	294	UNAVA	277	UNAV	71	60	106	UNAV	UNAV	UNAV	UNAVA	1340	1/30/09
Mycobacterium sp. Spyr1	UNAVAI	4888	79	UNAVA	992	UNAVA	99	UNAV	34	53	658	UNAV	UNAV	UNAV	UNAVA	2705	1/30/09
Aspergillus fumigatus Af293, chromosome 1	UNAVAI	1612	76	UNAVA	54	UNAVA	195	UNAV	44	4	518	UNAV	UNAV	UNAV	UNAVA	885	2/18/09
Zymomonas mobilis, draft genome, NCBI ID 11163	UNAVAI	1348	52	UNAVA	38	UNAVA	35	UNAV	34	8	59	UNAV	UNAV	UNAV	UNAVA	265	3/23/09
Porphyromonas gingivalis ATCC 33277	UNAVAI	2094	68	UNAVA	46	UNAVA	93	UNAV	0	23	908	UNAV	UNAV	UNAV	UNAVA	1552	7/29/09

Supplementary Data 2: Gene models in Meth before manual curation

ftp://ftp.jgi-psf.org/pub/JGI_data/apati/GenePRIMP_Supplementary_Data/meth_before_manual_curation.art

Supplementary Data 3: Gene models in Meth after manual curation

ftp://ftp.jgi-psf.org/pub/JGI_data/apati/GenePRIMP_Supplementary_Data/meth_after_manual_curation.art

Supplementary Data 4: Gene models in Myco before manual curation

ftp://ftp.jgi-psf.org/pub/JGI_data/apati/GenePRIMP_Supplementary_Data/myco_before_manual_curation.gb

Supplementary Data 5: Gene models in Myco after manual curation

ftp://ftp.jgi-psf.org/pub/JGI_data/apati/GenePRIMP_Supplementary_Data/myco_after_manual_curation.gb