Tripp et al. Standards in Genomic Sciences (2015) 10:45
DOI 10.1186/s40793-015-0034-9

Standards in Genomic Sciences

RESEARCH Open Access

Toward a standard in structural genome annotation for prokaryotes

H. James Tripp[1†], Granger Sutton[2], Owen White[3], Jennifer Wortman[4], Amrita Pati[1], Natalia Mikhailova[1], Galina Ovchinnikova[1], Samuel H. Payne[5], Nikos C. Kyrpides[1] and Natalia Ivanova[1*†]

Abstract

Background: In an effort to identify the best practice for finding genes in prokaryotic genomes and propose it as a standard for automated annotation pipelines, 1,004,576 peptides were collected from various publicly available resources, and were used as a basis to evaluate various gene-calling methods. The peptides came from 45 bacterial replicons with an average GC content from 31 % to 74 %, biased toward higher GC content genomes. Automated, manual, and semi-manual methods were used to tally errors in three widely used gene calling methods, as evidenced by peptides mapped outside the boundaries of called genes.

Results: We found that the consensus set of identical genes predicted by the three methods constitutes only about 70 % of the genes predicted by each individual method (with start and stop required to coincide). Peptide data was useful for evaluating some of the differences between gene callers, but not reliable enough to make the results conclusive, due to limitations inherent in any proteogenomic study.

Conclusions: A single, unambiguous, unanimous best practice did not emerge from this analysis, since the available proteomics data were not adequate to provide an objective measurement of differences in the accuracy between these methods. However, as a result of this study, software, reference data, and procedures have been better matched among participants, representing a step toward a much-needed standard. In the absence of sufficient amount of exprimental data to achieve a universal standard, our recommendation is that any of these methods can be used by the community, as long as a single method is employed across all datasets to be compared.

Background

As of July 13, 2013, more than a third of the 29,183 bacterial and archaeal genome sequencing projects listed in the Genomes On-line Database (GOLD) [1] are attributable to four major sequencing centers: DOE Joint Genome Institute (JGI, 4,250 projects), The Broad Institute (3,155 projects), J. Craig Venter Institute (JCVI, 1,976 projects), and Institute for Genome Sciences (IGS, 1,269 projects). Assuming an average of 3,000 gene predictions per genome for the 10,650 projects at these sequencing centers, an estimated 31,950,000 gene predictions will have been made by the completion of these projects. Given that each sequencing center has its own automated gene prediction pipeline, using software that has evolved separately over more than a decade, the question arises as to best current practices in structural genome annotation. In this context the phrase "structural gene annotation" refers only to finding the loci of protein-coding genes, not to annotating protein functions or predicting their 3D structure. Implementation of a single best practice would have the benefit of producing a single gene locus identifier for ease of cross-referencing in the scientific literature and for use by comparative genomics software [2]. A related motivation for this study was the need to consistently reannotate public genomes whose annotations are now more than a decade old.

Functional genomics data, such as RNA sequencing (RNA-Seq) and proteomics, provide a useful reference for evaluating and improving genome annotations [3–8]. A combination of the two is especially powerful, since RNA-seq data reveals transcript boundaries, whereas proteomics helps mapping translated sequences (coding sequences or CDSs). We found very few genomes where peptide data was available to confirm RNA-seq data, and

* Correspondence: nnivanova@lbl.gov
†Equal contributors
[1]DOE Joint Genome Institute, Walnut Creek, California, USA
Full list of author information is available at the end of the article

since all available prokaryotic gene finders predict translated products rather than transcript boundaries, we explored whether proteomics alone could serve as a tool to identify a best practice for updating gene calls in outdated genome annotations.

A test set of genomes with varying GC% was identified and the gene calls for GeneMarkS [9], Glimmer3 [10], and Prodigal [11], which are the three most popular *ab initio* methods, were obtained from RefSeq's public ftp site [12]. Peptides for most of the genomes were compiled from a PNNL website [8], with other data obtained from the PRIDE BioMart [13] and the publications of several independent research labs [6, 14, 15] (Additional file 1). We used the peptides to evaluate the accuracy of the GeneMarkS, Glimmer3, and Prodigal2.5 gene callers. In addition, we evaluated gene calling by the gene-finding post processor GenePRIMP [16] developed by JGI. Notably, the genome annotation versions of GeneMarkS, Glimmer3, Prodigal and GenePRIMP correspond to July 2013, when the work on this project was started. Based on the results of this work, reannotation of all public genomes integrated in IMG has begun at the JGI using the Prodigal based gene calling pipeline.

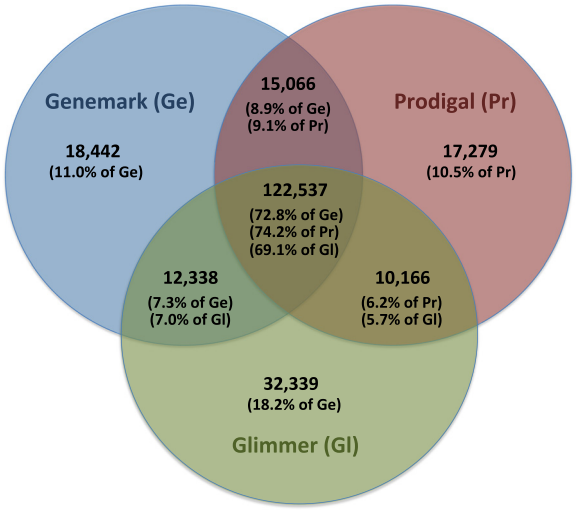

Fig. 1 Overlaps between the sets of identical genes predicted by the three ab initio gene callers for 52 genomes. Gene predictions by two gene callers coincide only if both of their start and stop codons are predicted to be in the same positions on the same strand. The numerator for the percentages reported on the diagram is the number of relevant calls, which appears above the percentage. The denominator for the percentages is the total number of calls made by the gene caller, whose abbreviation appears after the percentage. Ge, GeneMark; Gl, Glimmer; Pr, Prodigal

Results

Comparison of gene predictions

The consensus set of identical genes (same strand, start, and stop) predicted by the three methods for 45 replicons (Additional file 1) constitutes only about 67-73 % of the genes predicted by each individual method, depending on which gene caller is chosen to provide the total number of calls on which the percentage is based (Fig. 1). The consensus set of genes, which vary only in their start codon, constitutes 83-96 % of the genes predicted by each individual method (data not shown). With respect to unique predictions, Glimmer3 made nearly twice as many as Prodigal and GeneMarkS did. With respect to agreement between pairs of gene callers, Prodigal and GeneMarkS agreed most often while Prodigal and Glimmer3 agreed the least.

Peptide coverage of genes

Peptide coverage of predicted genes, which is to say the percentage of genes in the entire dataset that had at least one peptide mapping wholly inside of the gene, was on average approximately 40 % (data not shown). Total peptide support for gene calls, which is to say the total number of peptides that fell wholly inside of any gene prediction, was highest for Prodigal (1,000,574) and lowest for Glimmer3 (994,973) with GeneMarkS intermediate between the two (996,336).

Comparison of detectible errors

Among *ab initio* gene callers, Glimmer3 scored the most errors in total and in each error category, Prodigal scored the fewest, and GeneMarkS scored intermediate between the two (Fig. 2). The GenePRIMP post-processor scored fewer total errors than any of the *ab initio* gene callers.

In an effort to gain a more detailed understanding of the summary results, heat maps of wrong, short, and missed gene calls plotted for each replicon (Figs. 3, 4, 5). Total wrong gene calls showed this pattern with respect to number of errors: Glimmer3 > GeneMarkS > Prodigal > GenePRIMP (Fig. 3). This pattern followed the pattern of the overall results. That is, GenePRIMP lowered the errors made by Prodigal and the two methods combine for the lowest errors. Glimmer3 had the highest number of errors for wrong gene calls. The results for total short gene calls follows the same pattern (Fig. 4). The results for total missed calls shows a different pattern: Glimmer3 > GeneMarkS > GenePRIMP > Prodigal (Fig. 5). The reason for this difference is the presence of genes with interrupted translation frames, which GenePRIMP identifies as pseudogenes and which is further addressed in the discussion.

Discussion

A number of caveats must be kept in mind when attempting to estimate gene calling program performance from peptide data.

1. No proteomic experiment can guarantee expression of every gene in the genome.

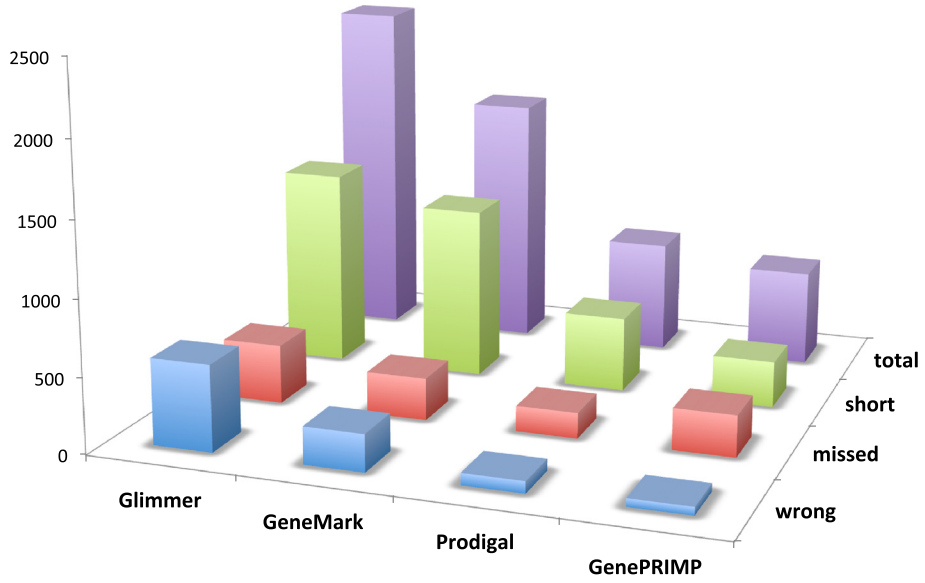

Fig. 2 Overview of all gene calling errors by gene calling method. The number of gene calling errors found in the entire data set, by type, are plotted by gene calling method

2. Signal peptides, are often removed from proteins, making it impossible to guarantee peptide data pertaining to the true start site of translation.
3. Some peptide sequences, particularly highly hydrophobic ones, are not amenable to detection by mass spectrometry.
4. The mapping of peptide mass spectra to genome sequence may be erroneous and thus presents an opportunity for false positives.
5. It is impossible to detect "too long" errors in gene start calling using peptide data, since an error correcting peptide will never appear upstream of a predicted start that is already upstream of the true start. It is important to recognize that it is possible for a lower "short" gene error rate to be offset by a higher "long" gene error rate, resulting in a better overall rate of calling correct gene starts. So the "short" gene error rate in itself is not an unbiased measurement of a gene finder's ability to choose correct gene start sites. However, considering that "short" gene errors prevent identification of functionally important conserved domains and motifs, and therefore can result in erroneous functional predictions, we report it here with this caveat in mind. In addition, we should point out that some genes have alternative translation initiation sites. This may have caused some spurious "short" errors, however all of the gene callers were under the same handicap in this regard.
6. It is impossible to detect false positive gene calls using peptide data, since peptides can only confirm gene calls; they cannot deny them.

In addition to these general caveats, it must also be reiterated that the genomes chosen for analysis are not a random, representative sample. Therefore, the results presented here must be considered an estimate of gene calling performance detectable with proteomics, not a definitive and absolute measurement of true gene calling performance.

At the same time, there is no definitive measurement of true gene calling performance against a randomly chosen, fully representative set of genomes. The biological knowledge to force expression of every protein in every genome does not exist, nor do high throughput biochemical methods for detecting every amino acid residue in every translation product in a cell, even if such knowledge of expression were available. The expression rate for this study, as measured by the peptide coverage reported above, averaged less than half (~40 %), but there is no reason to assume that this sample is systematically biased for or against any particular gene caller. Also, while it is true that peptides cannot detect false positive gene calls, statistical observations can give evidence of false positives: Glimmer3 made twice as many unique gene calls as Prodigal or GeneMarkS, but had the fewest number of confirming peptides. This does not prove that it makes more false positive predictions than the other gene callers; it simply offers some evidence that it might. High throughput proteomic data is the only option available for performing a wide survey of gene calling accuracy for thousands of genes in dozens of genomes. Use of high throughput proteomics is therefore an operational necessity if one wishes to perform a survey of gene-

Organism	Accession	%GC	GP	Pr	GM	GI
Prochlorococcus marinus CCMP1986	NC_005072	30.8	0	0	0	2
Bacillus cereus ATCC 14579	NC_004722	35.3	0	0	0	0
Bacillus anthracis str. Sterne	NC_005945	35.4	0	0	0	2
Cyanothece sp. ATCC 51142	NC_010546	38.0	0	0	0	4
Streptococcus pyogenes M1 GAS	NC_002737	38.5	0	0	0	0
Clostridium thermocellum ATCC 27405	NC_009012	39.0	0	0	1	1
Methanosarcina barkeri Fusaro, DSM 804	NC_007355	39.0	0	0	1	7
Anabaena variabilis ATCC 29413	NC_007413	41.4	1	1	2	12
Anaplasma phagocytophilum HZ	NC_007797	41.6	0	0	0	0
Yersinia pestis KIM10+	NC_004088	47.7	0	0	0	6
Synechocystis sp. PCC 6803	NC_000911	47.4	0	0	6	14
Synechococcus sp. PCC 7002	NC_010475	49.2	1	1	6	13
Escherichia coli str. K-12 substr. MG1655	NC_000913	50.8	1	0	2	9
Cryptobacterium curtum DSM 15641	NC_013170	50.9	1	1	1	7
Thermococcus gammatolerans EJ3 *	NC_012804	54.0	2	2	2	2
Geobacter uraniireducens Rf4	NC_009483	54.2	3	6	9	31
Pelobacter carbinolicus DSM 2380	NC_007498	55.1	0	0	1	5
Chlorobium tepidum TLS	NC_002932	56.5	0	3	10	13
Chloroflexus aurantiacus J-10-fl	NC_010175	56.7	2	0	3	6
Desulfovibrio alaskensis G20	NC_007519	57.8	1	5	12	17
Geobacter metallireducens GS-15	NC_007517	59.5	5	7	17	23
Syntrophobacter fumaroxidans MPOB	NC_008554	59.9	1	0	4	11
Halogeometricum borinquense PR3, DSM 11551	NC_014729	60.0	0	2	1	3
Slackia heliotrinireducens DSM 20476	NC_013165	60.2	0	0	2	9
Roseiflexus castenholzii DSM 13941	NC_009767	60.7	0	0	9	17
Geobacter sulfurreducens PCA	NC_002939	60.9	0	0	8	7
Halorhabdus utahensis AX-2, DSM 12940	NC_013158	63.0	5	5	13	4
Deinococcus deserti VCD115	NC_012526	63.0	2	4	0	1
Desulfovibrio vulgaris str. Hildenborough	NC_002937	63.2	2	4	13	22
Arthrobacter sp. FB24	NC_008541	65.4	3	5	14	42
Mycobacterium tuberculosis H37Rv	NC_000962	65.6	0	2	10	21
Rhodobacter capsulatus SB 1003	NC_014034	66.6	1	1	13	21
Deinococcus radiodurans R1	NC_001263	66.6	5	4	16	18
Saccharomonospora viridis DSM 43017	NC_013159	67.3	1	1	3	20
Mycobacterium smegmatis str. MC2 155	NC_008596	67.4	0	0	0	3
Stackebrandtia nassauensis DSM 44728	NC_013947	68.1	7	8	19	22
Rhodobacter sphaeroides 2.4.1	NC_007494	68.8	0	1	0	3
Rhodobacter sphaeroides 2.4.1	NC_007494	68.8	10	10	11	22
Brachybacterium faecium DSM 4810	NC_013172	72.0	1	2	9	18
Thermobispora bispora DSM 43833	NC_014165	72.4	1	1	4	9
Xylanimonas cellulosilytica DSM 15894	NC_013530	72.5	1	2	8	37
Nocardiopsis dassonvillei DSM 43111	NC_014210	72.7	1	1	2	12
Actinosynnema mirum DSM 43827	NC_013093	73.7	1	1	6	18
Kineococcus radiotolerans SRS30216	NC_009664	74.2	0	3	21	50
Cellulomonas flavigena DSM 20109	NC_014151	74.3	2	4	13	37
Total "Wrong" Errors			**61**	**87**	**272**	**601**

Fig. 3 Total wrongly predicted (annotated) genes. GP, GenePRIMP; Pr, Prodigal; GM, GeneMarkS; GI, Glimmer3

calling methods in preparation for a task such as reannotating all public genomes.

An important parameter affecting gene predictions made by *ab initio* gene callers is minimum gene length. Other things being equal, a shorter minimum gene length yields more candidate ORFs, which can result in a larger number of genes called. A biologically meaningful minimum gene length is 39 nucleotides (nt), which is the length of the PatS peptide, the shortest CDS yet detected [17]. However, such a short length generates so many spurious candidate ORFs that it is not recommended by designers of *ab initio* gene callers. The default minimum gene lengths recommended by program designers are: 90 nt for Prodigal, 81 nt for GeneMarkS, and 120 nt for Glimmer. These defaults were suggested by the developers of the corresponding tools to ensure their optimal performance. Selecting any other minimum gene length cutoff than 39 nt cannot be biologically justified but will undoubtedly result in poor performance; furthermore, it is likely to bias the analysis against one or another gene finder. For these reasons we chose to proceed with default cutoffs.

Turning to an analysis of the data regarding the three *ab initio* gene callers, it appears that Glimmer3's "aggressive" algorithm for finding novel coding regions makes it prone to errors detectable with proteomics, while Prodigal's design objective of eliminating false positives while retaining sensitivity makes it the least prone to such errors. The version of GeneMarkS tested, which now has been improved but not yet released, produced intermediate results. It is possible that the "aggressive" gene calling of Glimmer might be appropriate for a different set of genomes from novel single cell organisms. It may also be that the careful modeling of non-coding regions done by GeneMarkS (Mark Borodovsky, personal communication) may avoid more false positives than Prodigal in genomes with exceptionally low coding percentage and low GC content.

Organism	Accession	%GC	GP	Pr	GM	Gl
Prochlorococcus marinus CCMP1986	NC_005072	30.8	1	2	4	16
Bacillus cereus ATCC 14579	NC_004722	35.3	1	1	1	1
Bacillus anthracis str. Sterne	NC_005945	35.4	1	2	2	13
Cyanothece sp. ATCC 51142	NC_010546	38.0	2	2	9	20
Streptococcus pyogenes M1 GAS	NC_002737	38.5	0	0	0	0
Clostridium thermocellum ATCC 27405	NC_009012	39.0	0	0	0	0
Methanosarcina barkeri Fusaro, DSM 804	NC_007355	39.0	7	16	20	33
Anabaena variabilis ATCC 29413	NC_007413	41.4	5	6	25	39
Anaplasma phagocytophilum HZ	NC_007797	41.6	2	2	1	6
Yersinia pestis KIM10+	NC_004088	47.7	1	3	8	20
Synechocystis sp. PCC 6803	NC_000911	47.4	2	2	14	35
Synechococcus sp. PCC 7002	NC_010475	49.2	5	10	35	36
Escherichia coli str. K-12 substr. MG1655	NC_000913	50.8	2	2	12	31
Cryptobacterium curtum DSM 15641	NC_013170	50.9	5	6	10	18
Thermococcus gammatolerans EJ3 *	NC_012804	54.0	2	4	4	3
Geobacter uraniireducens Rf4	NC_009483	54.2	7	15	33	54
Pelobacter carbinolicus DSM 2380	NC_007498	55.1	0	0	2	5
Chlorobium tepidum TLS	NC_002932	56.5	7	14	43	45
Chloroflexus aurantiacus J-10-fl	NC_010175	56.7	3	6	22	30
Desulfovibrio alaskensis G20	NC_007519	57.8	11	17	41	43
Geobacter metallireducens GS-15	NC_007517	59.5	5	12	35	43
Syntrophobacter fumaroxidans MPOB	NC_008554	59.9	1	1	7	14
Halogeometricum borinquense PR3, DSM 11551	NC_014729	60.0	0	0	8	22
Slackia heliotrinireducens DSM 20476	NC_013165	60.2	2	7	15	17
Roseiflexus castenholzii DSM 13941	NC_009767	60.7	6	7	18	22
Geobacter sulfurreducens PCA	NC_002939	60.9	4	10	32	32
Halorhabdus utahensis AX-2, DSM 12940	NC_013158	63.0	5	6	13	11
Deinococcus deserti VCD115	NC_012526	63.0	1	1	8	12
Desulfovibrio vulgaris str. Hildenborough	NC_002937	63.2	7	12	33	53
Arthrobacter sp. FB24	NC_008541	65.4	29	43	82	97
Mycobacterium tuberculosis H37Rv	NC_000962	65.6	24	26	49	78
Rhodobacter capsulatus SB 1003	NC_014034	66.6	22	24	54	47
Deinococcus radiodurans R1	NC_001263	66.6	26	38	53	57
Saccharomonospora viridis DSM 43017	NC_013159	67.3	7	17	35	49
Mycobacterium smegmatis str. MC2 155	NC_008596	67.4	0	6	27	26
Stackebrandtia nassauensis DSM 44728	NC_013947	68.1	10	21	48	38
Rhodobacter sphaeroides 2.4.1	NC_007494	68.8	6	6	17	14
Rhodobacter sphaeroides 2.4.1	NC_007494	68.8	21	30	68	61
Brachybacterium faecium DSM 4810	NC_013172	72.0	10	16	31	31
Thermobispora bispora DSM 43833	NC_014165	72.4	1	5	28	16
Xylanimonas cellulosilytica DSM 15894	NC_013530	72.5	20	30	56	53
Nocardiopsis dassonvillei DSM 43111	NC_014210	72.7	4	11	51	34
Actinosynnema mirum DSM 43827	NC_013093	73.7	18	17	50	39
Kineococcus radiotolerans SRS30216	NC_009664	74.2	20	59	86	77
Cellulomonas flavigena DSM 20109	NC_014151	74.3	14	22	35	37
Total "Short" Errors			**327**	**537**	**1225**	**1428**

Fig. 4 Genes with starts predicted downstream from detected starts (as indicated by proteomics). GP, GenePRIMP; Pr, Prodigal; GM, GeneMarkS; Gl, Glimmer3

We explored the hypothesis that all gene callers might show poorer performance at high GC because they contain a higher frequency of alternate start codons and a lower frequency of stop codons. Our analysis uncovered two biases in our dataset that prevented rigorous exploration of this hypothesis: increased genome size with increasing GC content (Fig. 6a) and increased number of peptides with increasing GC content (Fig. 6b). Larger genomes are likely to have more total errors, and genomes with more peptides are more likely to have detectable errors. Future research might call for additional proteomics datasets generated with specific purpose of improving prokaryotic structural annotation by carefully selecting an unbiased set of genomes that will shed more light on the specific reasons for differences in performance of gene callers across genomes and for specific genomes.

Although post processing by GenePRIMP generally improved upon Prodigal's predictions, the number of missed genes was higher in GenePRIMP annotation due to the presence of pseudogenes. The issue of pseudogene calling does not arise with *ab initio* gene callers; they simply search for coding domains and make no attempt to analyze whether adjacent coding domains are part of a pseudogene. With regard to post processing by GenePRIMP, it must be noted that it does not automatically assume that all genes with frame disruptions are pseudogenes. Instead, it considers the number of frame disruptions (frameshifts and/or stop codons) and the length of the gene as compared to its homologs, and marks as pseudogenes only those with multiple frame disruptions and/or severe truncations. Furthermore, GenePRIMP retains the coordinates of all fragments of disrupted CDSs, even when they are annotated as pseudogenes. Frameshifted genes without a "pseudogene" tag that had confirming peptides were considered good calls despite their frameshifts, since the confirming peptides indicate that not calling the gene a pseudogene was a correct

Organism	Accession	%GC	Pr	GP	GM	GI
Prochlorococcus marinus CCMP1986	NC_005072	30.8	3	4	2	3
Bacillus cereus ATCC 14579	NC_004722	35.3	0	4	1	0
Bacillus anthracis str. Sterne	NC_005945	35.4	0	3	0	1
Cyanothece sp. ATCC 51142	NC_010546	38.0	2	6	2	1
Streptococcus pyogenes M1 GAS	NC_002737	38.5	1	2	2	1
Clostridium thermocellum ATCC 27405	NC_009012	39.0	1	0	1	2
Methanosarcina barkeri Fusaro, DSM 804	NC_007355	39.0	4	13	0	4
Anabaena variabilis ATCC 29413	NC_007413	41.4	3	9	1	4
Anaplasma phagocytophilum HZ	NC_007797	41.6	0	2	0	0
Yersinia pestis KIM10+	NC_004088	47.7	0	1	0	1
Synechocystis sp. PCC 6803	NC_000911	47.4	0	7	7	3
Synechococcus sp. PCC 7002	NC_010475	49.2	2	15	7	15
Escherichia coli str. K-12 substr. MG1655	NC_000913	50.8	1	4	3	8
Cryptobacterium curtum DSM 15641	NC_013170	50.9	2	3	3	6
Thermococcus gammatolerans EJ3 *	NC_012804	54.0	1	7	1	6
Geobacter uraniireducens Rf4	NC_009483	54.2	14	12	12	16
Pelobacter carbinolicus DSM 2380	NC_007498	55.1	0	2	1	4
Chlorobium tepidum TLS	NC_002932	56.5	9	10	6	9
Chloroflexus aurantiacus J-10-fl	NC_010175	56.7	2	11	6	4
Desulfovibrio alaskensis G20	NC_007519	57.8	5	13	10	18
Geobacter metallireducens GS-15	NC_007517	59.5	12	15	17	18
Syntrophobacter fumaroxidans MPOB	NC_008554	59.9	2	12	6	8
Halogeometricum borinquense PR3, DSM 11551	NC_014729	60.0	2	0	3	3
Slackia heliotrinireducens DSM 20476	NC_013165	60.2	0	1	2	0
Roseiflexus castenholzii DSM 13941	NC_009767	60.7	0	8	10	14
Geobacter sulfurreducens PCA	NC_002939	60.9	1	8	15	11
Halorhabdus utahensis AX-2, DSM 12940	NC_013158	63.0	3	9	18	12
Deinococcus deserti VCD115	NC_012526	63.0	5	5	8	7
Desulfovibrio vulgaris str. Hildenborough	NC_002937	63.2	10	14	15	18
Arthrobacter sp. FB24	NC_008541	65.4	4	5	14	15
Mycobacterium tuberculosis H37Rv	NC_000962	65.6	5	5	11	15
Rhodobacter capsulatus SB 1003	NC_014034	66.6	11	13	17	22
Deinococcus radiodurans R1	NC_001263	66.6	14	14	24	22
Saccharomonospora viridis DSM 43017	NC_013159	67.3	2	2	3	1
Mycobacterium smegmatis str. MC2 155	NC_008596	67.4	4	0	5	6
Stackebrandtia nassauensis DSM 44728	NC_013947	68.1	11	13	18	21
Rhodobacter sphaeroides 2.4.1	NC_007494	68.8	7	5	8	7
Rhodobacter sphaeroides 2.4.1	NC_007494	68.8	14	13	23	22
Brachybacterium faecium DSM 4810	NC_013172	72.0	5	5	5	8
Thermobispora bispora DSM 43833	NC_014165	72.4	2	4	4	3
Xylanimonas cellulosilytica DSM 15894	NC_013530	72.5	3	5	5	13
Nocardiopsis dassonvillei DSM 43111	NC_014210	72.7	1	5	1	1
Actinosynnema mirum DSM 43827	NC_013093	73.7	0	5	3	4
Kineococcus radiotolerans SRS30216	NC_009664	74.2	5	11	7	38
Cellulomonas flavigena DSM 20109	NC_014151	74.3	7	8	11	35
Total "Missed" Errors			**180**	**313**	**318**	**430**

Fig. 5 Genes missed by gene prediction (annotation) methods. Pr, Prodigal; GP, GenePRIMP; GM, GeneMarkS; GI, Glimmer3

call even though frameshifts were present. On the other hand, GenePRIMP pseudogene calls with confirming peptides were scored as "missed," since the confirming peptides indicate that the gene in question is a real gene, not a pseudogene, despite the fact that it may have multiple frameshifts due to sequencing and assembly errors (Fig. 7). This explains why missed gene errors for GenePRIMP are higher than Prodigal's.

A similar situation occurs when sequencing error introduces an interrupted gene. GenePRIMP attempts to detect these instances and joins coding domains it deems to have likely been interrupted by the introduction of a spurious stop codon due to sequencing error. The partial coding domains are annotated as "exons," even though they are not pieces of a gene interrupted by introns (as in eukaryotes), but rather pieces of a gene with spurious interruptions introduced by sequencing errors that result in frameshifts and internal stop codons. The missed gene data suggests that the GenePRIMP algorithm for detecting and annotating sequencing error is reliable and appropriate, albeit with a novel interpretation of "exon." For similar reasons, NCBI has recently changed the guidelines for annotation of interrupted genes in RefSeq genomes (personal communication). Interrupted genes are annotated as partial coding regions if their translated protein products have significant similarity to full-length proteins in closely related genomes.

Conclusions

Proteomics is a valuable aid to evaluating and improving gene-calling programs. When applied to 45 replicons of interest to the participants of this study, a combination of *ab initio* gene calling by Prodigal followed by GenePRIMP post processing had a lower estimated, operational error rate than GeneMarkS followed by Glimmer3. We have also compared these data against the RefSeq pipeline (version available in Spring 2013) and the results showed that the its overall performance

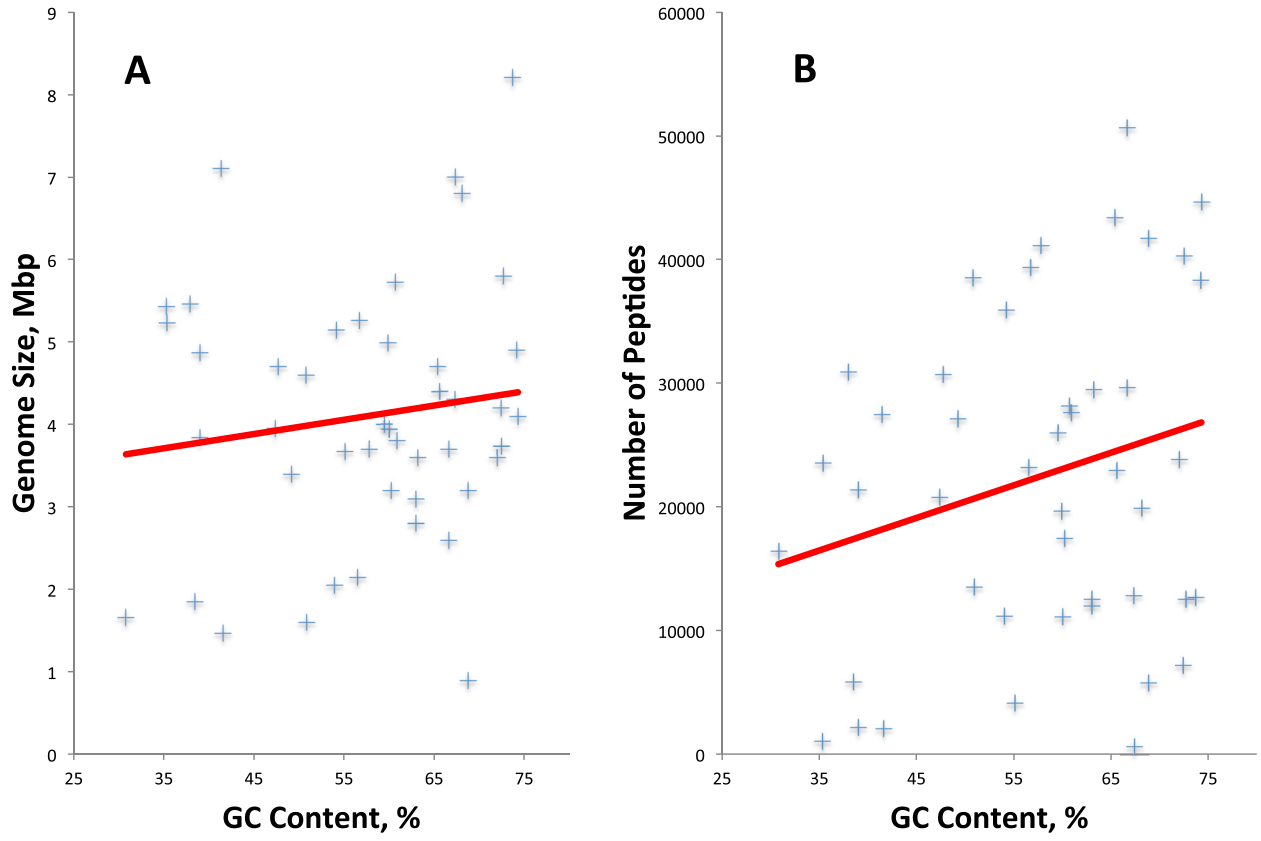

Fig. 6 Bias toward increased genome length and number of peptides with increased GC content. **a**. The genome length in Mbp is plotted against GC content of replicons used. **b**. The number of peptides is plotted against GC content of replicons used

was between that of GeneMark and Glimmer (data not shown). Nonetheless, due to inherent biologically-based limitations, we cannot conclude that proteomics alone should be used to define a best practice as the basis for a general standard in prokaryotic structural genome annotation; this must wait for better tools and expanded datasets that cover more taxonomic groups without biases in GC content, genome length, and gene expression levels. Some participants have already improved their pipelines, especially gene data models and reference databases, with the goal of one day achieving a much needed standard. Moving forward, a consensus approach, employing multiple gene callers and additional forms of expression verification such as RNA-seq, should be also explored as the possible basis for a standard in prokaryotic structural genome annotation.

Methods

Selection of genomes of interest

The genomes of interest had at least one of these characteristics:

1. The annotation was thought to be in need of updating.

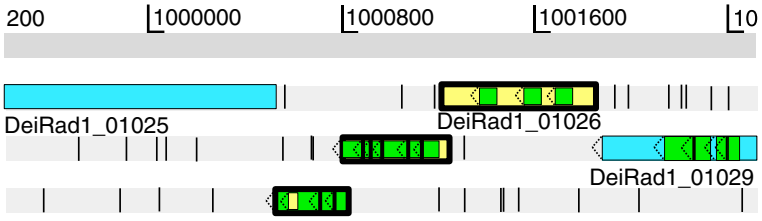

Fig. 7 Artemis visualization of peptides refuting incorrect pseudogene call. The three boxes with thick black outlines and yellow backgrounds are CDS fragments of a single gene (DeiRad1_01026) disrupted by two frameshifts. The green boxes represent detected peptides

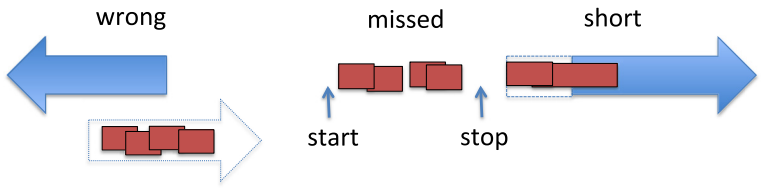

Fig. 8 Schematic representation of scoring errors in gene calling. Right and left pointing arrows indicate genes called on positive and negative genome strand respectively. Boxes represent peptides detected by proteomics. Dashed contours show the extension of a gene or missed gene implied by peptide data

2. The organism was well studied; preferably a type strain whose annotation had been heavily curated.
3. The genome added taxonomic diversity to the dataset.

Because of the genome selection criteria, the sample set is diverse and relevant to the participants, but it cannot be considered a representative random sample of genomes in nature or in public databases. The average GC content was 57.7 %, with a range of 30.8 % to 74.3 %. The average genome size was 4.3 Mbp with a range of 1.5 Mbp to 8.2 Mbp.

Collection of data and loading to MySQL warehouse

A mySQL warehouse of proteogenomic information was created by acquiring, transforming, and loading publically available gene calls and proteomic data. The data sources for gene calls and peptides are shown in Additional file 1 and are available at http://portal.nersc.gov/dna/microbial/prokpubs/SIGS_proteogenomics/. GeneMarkS, Glimmer3, and Prodigal-2.5 in .gff format were downloaded from RefSeq public ftp site at the time of this study (Spring, 2013). These gene call coordinates were extracted from the .gff files and loaded into the warehouse. It must be noted that as a result of this study, development of a new version of GeneMarkS has started (GeneMarkS-2), however, predictions made by the new version have not been used in this study. The PNNL peptide data was also provided in .gff format, with mapping to its associated GenBank nucleotide sequence, allowing easy extraction and loading into MySQL. The non-PNNL peptide data was often not provided in .gff format, and sometimes did not have end coordinates. However, it always included individual peptide sequences, allowing each peptide to be mapped to its coordinates in the corresponding GenBank fasta file. Mapping was accomplished using a Perl script that searched for exact, unique matches in one of the six translation frames of the corresponding nucleotide sequence for the peptide. Short peptides that could not be mapped unambiguously were discarded. Unambigously mapped peptides were loaded into MySQL.

Identification and analysis of peptides conflicting with gene calls

A SQL script identified peptides whose boundaries were partially or fully outside of a gene call. A Perl script produced a .gff file of conflicting peptides for each gene calling method. This .gff file of peptides and its corresponding nucleotide sequence were loaded into Artemis. A JGI Quality Assurance Analyst scored the gene associated with the conflicting peptide as wrong, missed, or short (Fig. 8). When conflicting peptides lay in the wrong reading frame relative to the gene call, the gene call was scored as "wrong." When at least two conflicting peptides extended upstream of the predicted start site, the gene call was scored "short." When conflicting peptides were in a region without any gene calls the implied gene was scored as "missed." The false discovery rate (FDR) for the majority of the data was reported as 0.3 % [8]. In order to further reduce the rate of false positives, we required that each missed, wrong or short gene was detected by at least 2 non-redundant peptides, which would reduce the false positive rate to 0.09 %. All missed, wrong and short genes were additionally verified by BLASTp with an e-value of 1.0e-05 and by conserved motif and domain analysis.

Additional file

Additional file 1: Replicons Used in This Study. This is an Excel file listing the GenBank accession numbers of the 45 replicons used in this study, along with the number of peptides, the RefSeq source for gene calls, and the file name and publication reference for the peptide sources.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NCK conceived the study, oversaw the project, and analyzed data. NI oversaw collection, analysis, and quality assurance of the data. HJT collected and pre-processed data, and also wrote the manuscript. AP loaded data into a warehouse and produced outputs for analysis. NM and GO did quality assurance on data outputs and tallied results. All authors read and approved the manuscript.

Acknowledgements
The authors wish to thank Doug Hyatt of Oak Ridge National Laboratory for a detailed description of the algorithms employed by Prodigal as well as Mark Borodovsky (Georgia Tech) and Tatiana Tatusova (NCBI) for helpful comments and discussion. 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.

Author details
[1]DOE Joint Genome Institute, Walnut Creek, California, USA. [2]J. Craig Venter Institute, Rockville, MD, USA. [3]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA. [4]Broad Institute, Cambridge, MA, USA. [5]Pacific Northwest National Laboratory, Richland, WA, USA.

Received: 18 December 2014 Accepted: 1 July 2015
Published online: 25 July 2015

References

1. Reddy TB, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, et al. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 2014, doi: 10.1093/nar/gku950.
2. Kyrpides NC. Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. Nat Biotechnol. 2009;27(7):627–32.
3. Tanner S, Shen Z, Ng J, Florea L, Guigó R, Briggs SP, et al. Improving gene annotation using peptide mass spectrometry. Genome Res. 2007;17:231–9.
4. de Souza GA, Softeland T, Koehler CJ, Thiede B, Wiker HG. Validating divergent ORF annotation of the *Mycobacterium leprae* genome through a full translation data set and peptide identification by tandem mass spectrometry. Proteomics. 2009;9:3233–43.
5. Zivanovic Y, Armengaud J, Lagorce A, Leplat C, Guérin P, Dutertre M, et al. Genome analysis and genome-wide proteomics of *Thermococcus gammatolerans*, the most radioresistant organism known amongst the Archaea. Genome Biol. 2009;10:R70.
6. Baudet M, Ortet P, Gaillard JC, Fernandez B, Guérin P, Enjalbal C, et al. Proteomics-based refinement of *Deinococcus deserti* genome annotation reveals an unwonted use of non-canonical translation initiation codons. Molecular Cellular proteomics: MCP. 2010;9:415–26.
7. Castellana N, Bafna V. Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteome. 2010;73:2124–35.
8. Venter E, Smith RD, Payne SH. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One. 2011;6, e27587.
9. Borodovsky M, Lomsadze A. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. Curr Protoc Microbiol. 2014;32(1E):7.
10. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007;23(6):673–9.
11. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ, et al. Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
12. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, et al. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res. 2014;42:D553–559.
13. Vizcaíno JA, Côté RG, Csordas A, Dianes JA, Fabregat A, Foster JM, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. 2013;42013, 1:D1063–1069. doi:D1063.
14. Gallien S, Perrodou E, Carapito C, Deshayes C, Reyrat JM, Van Dorsselaer A, et al. Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res. 2009;19:128–35.
15. Clair G, Roussi S, Armengaud J, Duport C. Expanding the known repertoire of virulence factors produced by Bacillus cereus through early secretome profiling in three redox conditions. Molecular Cellular proteomics: MCP. 2010;9:1486–98.
16. Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, et al. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. Nat Methods. 2010;7:455–7.
17. Yoon HS, Golden JW. Heterocyst pattern formation controlled by a diffusible peptide. Science. 1998;282:935–8.