Published in final edited form as:

Curr Microbiol. 2018 December ; 75(12): 1642–1648. doi:10.1007/s00284-018-1572-3.

Achieving accurate sequence and annotation data for *Caulobacter vibrioides* CB13

Louis Berrios and Bert Ely

Department of Biological Sciences, University of South Carolina, Columbia, South Carolina, United States of America 29208

Abstract

Annotated sequence data is instrumental in nearly all realms of biology. However, the advent of next-generation sequencing has rapidly facilitated an imbalance between accurate sequence data and accurate annotation data. To increase the annotation accuracy of the *Caulobacter vibrioides* CB13b1a (CB13) genome, we compared the PGAP and RAST annotations of the CB13 genome. A total of 64 unique genes were identified in the PGAP annotation that were either completely or partially absent in the RAST annotation, and a total of 16 genes were identified in the RAST annotation that were not included in the PGAP annotation. Moreover, PGAP identified 73 frameshifted genes and 22 genes with an internal stop. In contrast, RAST annotated the larger segment of these frameshifted genes without indicating a change in reading frame may have occurred. The RAST annotation did not include any genes with internal stop codons, since it chose start codons that were after the internal stop. To confirm the discrepancies between the two annotations and verify the accuracy of the CB13 genome sequence data, we re-sequenced and re-annotated the entire genome and obtained an identical sequence, except in a small number of homopolymer regions. A genome sequence comparison between the two versions allowed us to determine the correct number of bases in each homopolymer region, which eliminated frameshifts for 31 genes annotated as frameshifted genes and removed 24 pseudogenes from the PGAP annotation. Both annotation systems correctly identified genes that were missed by the other system. In addition, PGAP identified conserved gene fragments that represented the beginning of genes, but it employed no corrective method to adjust the reading frame of frameshifted gene or the start sites of genes harboring an internal stop codon. In doing so, the PGAP annotation identified a large number of pseudogenes, which may reflect evolutionary history but likely do not produce gene products. These results demonstrate that re-sequencing and annotation comparisons can be used to increase the accuracy of genomic data and the corresponding gene annotation.

Introduction

Since the first complete bacterial genome sequence was published in 1995, the science of bacteriology has changed remarkably. With the advent of next-generation sequencing technologies, the nucleotide sequence of an entire bacterial genome can be determined in hours, as compared to the months it may have taken in the mid 90's. Particularly, third-generation sequencing such as that offered by Pacific Biosciences (PacBio) offers a fast

and accurate method to completely sequence a bacterial chromosome and assemble it into a single contig [14, 15].

In order to gain significant insight into the genome's genetic functionalities, the genome must be annotated. Generally, most annotation systems work similarly but each has its strengths and weaknesses. For example, annotation systems omit genes, annotate incorrect genes, identify frameshifts and include internal stops [9]. Two commonly used annotation platforms are the NCBI GenBank Prokaryotic Genome Annotation Pipeline (PGAP) and the Rapid Annotations using Subsystems Technology (RAST). PGAP utilizes a pan-genome approach to protein annotation with pan-genome proteins defined for each specific clade as well as a two-pass approach designed to detect frameshifted genes and pseudogenes [11, 16]. In contrast, RAST identifies approximately 30 nearest phylogenetic neighbors by comparing *ab initio* GLIMMER3 gene-candidates with a set of universal proteins to subsequently perform subsystem analyses and metabolic reconstructions [1, 10]. Gene fragments that potentially contain shifts in the reading frame are detected by a comparison with the template genes in the 30 nearest neighbors' phylogeny. The greatest difference between the two systems, however, is that PGAP calculates a set of alignment-based hints for protein-coding regions and non-protein-coding regions prior to executing *ab initio* gene prediction. Consequently, this built-in algorithm can cause PGAP annotations to accrue an overabundance of annotated genes, particularly in frameshifted genes. Thus, two major points can be deduced from the general analytical parameters among these systems: 1) PGAP may over-represent the quantity of frameshifted genes in comparison to RAST because of the relative sensitivity in its annotation method, and 2) PGAP should be more sensitive in its detection of novel genes but will produce more pseudogenes in the process. Although, both of these annotation programs correctly identify more 95% of the genes present in a bacterial genome, a comparison of the annotations produced by the two systems could be employed to obtain a more accurate annotation.

Ideally, an annotated sequence that has been experimentally verified such as the *C. crescentus* (*vibrioides*) NA1000 genome sequence [2, 6, 8, 13] would provide the most accurate representation of an annotated genome, but this approach is labor-intensive and expensive. Alternatively, annotation visualization programs such as Artemis [12] can be implemented to visualize differences in start sites, stop sites, reading frame position, coding region GC content, frameshifts, and internal stops, which all are commonplace annotation problems for genomes deposited in NCBI GenBank [6, 7]. Likewise the genome annotation alignment tool Mauve can be utilized to visualize the differences in gene loci between two or more annotated sequences [5]. Using Artemis and Mauve, we compared the PGAP and RAST annotations of the *C. vibrioides* CB13b1a (CB13) genome using the NA1000 genome as a reference. Additionally, the CB13 genome was sequenced a second time using repeated PacBio sequencing technology to verify the results of the annotation analysis and increase both the sequence and the annotation accuracy.

Methods

Annotation analysis

The annotated nucleotide sequence of the CB13b1a (CB13) genome was submitted to both GenBank (<http://www.ncbi.nlm.nih.gov>) and RAST (<http://rast.nmpdr.org/rast.cgi>) for annotation. The two CB13 annotations were aligned in Mauve [4, 5] to visualize gene annotation inconsistencies. When a gene annotation inconsistency was observed by manual inspection, the corresponding region of each genome was visualized in Artemis [12]. The GC frameplot feature of Artemis was then used to predict whether a particular region coded for a protein and to visualize the location of the start codon since *C. crescentus* coding regions generally have very high third position GC content [6]. In addition, the differing gene annotations were matched with known amino acid sequences in GenBank using BLASTp. If the E-value for a selected gene was less than the cutoff of 10^{-5} , with a query coverage of at least 60%, then the gene was assumed to be a potential gene. The experimentally verified conspecific *C. crescentus* NA1000 genome annotation (GenBank RefSeq NC_011916.1) was considered to be an accurate reference for comparison to the CB13 annotations.

Sequence analysis and modification

The complete genome of CB13 was sequenced using a PacBio RSII single-molecule sequencer and retrieved as a single contig that contained the entire genome. After the initial annotation comparison, the CB13 genome was sequenced a second time using the same PacBio technology. The nucleotide sequence of the genome was again obtained as a single contig. This second independent genome sequence was then aligned with the previous CB13 sequence using the progressiveMauve aligner [4]. A gap file containing all base regions that differed between the two sequences was exported and used to pinpoint the inconsistencies, all of which were located in GC rich homopolymer regions (Table S1). Since we have shown that this type of homopolymer sequencing error occurs randomly, and always is one base pair fewer than the actual homopolymer sequence (Ely et al., manuscript in preparation), we added an appropriate base pair to the original CB13 nucleotide sequence wherever the re-sequenced genome contained an extra base pair. After correcting the CB13 nucleotide sequence, we submitted the revised sequence to GenBank and RAST to be annotated a second time. The resulting annotations were then analyzed as previously described and compared to the first set of annotations.

Results and Discussion

A comparison of the initial PGAP and RAST annotations of the CB13 genome showed 90 discrepancies in the total number of genes that were annotated. However, 10 of these discrepant genes (additional genes annotated by PGAP) were smaller than 20 amino acids in length, were not present in the NA1000 reference genome annotation, and produced BLASTp hypothetical protein homologies with <60% query coverage and $>10^{-5}$ E-values, so they are probably not expressed genes (Table 1). Of the remaining discrepancies, 64 genes with statistically significant database matches were present in the PGAP annotation but were excluded from the RAST annotation, and 16 genes with statistically significant

database matches were present in the RAST annotation but were excluded from the PGAP annotation (Figure 1 and Table S2). For example, PGAP annotated gene CA606_07250, whereas RAST did not annotate this gene (Figure 2). Further, 51 of the 64 genes identified solely by PGAP had significant BLASTp matches to the well-studied NA1000 genome indicating that they were probably expressed genes. The remaining 13 genes also had significant BLASTp matches, with seven genes matching *Caulobacter* genomes other than NA1000 and six genes matching non-*Caulobacter* genera (Table S2). These failures to match the NA1000 genome were not a concern, because in each case the corresponding region of the genome was not present in the NA1000 genome. Thus, all 13 of these genes were considered expressed genes as well. Similarly, nine of the 16 genes included in the RAST annotation but excluded in the PGAP annotation were considered expressed genes since they had significant BLASTp matches to the well-studied NA1000 genome. Six of the remaining seven RAST genes had significant BLASTp matches to other *Caulobacter* genomes, while the remaining gene significantly matched non-*Caulobacter* genera in the database. Thus, PGAP failed to identify 16 verified genes that were accurately annotated by RAST, and RAST failed to identify 64 verified genes that were accurately annotated by PGAP.

In addition to differences in the number of genes annotated, PGAP identified 73 frameshifted genes (Table S3). Of the 73, RAST annotated 72 of these genes using a reading frame that showed no sign of a frameshift mutation, and the remaining gene was not annotated as a gene by RAST. The differences seem to be derived from the differences in the way that the two programs annotate genes, where the coding region is split into two reading frames. It appears that PGAP annotates a gene that is designated as frameshifted based on the reading frame at the beginning of the gene and then continues in the same reading frame until it reaches the end of the coding region, ignoring both changes in the reading frame and stop codons. In contrast, RAST annotates only the reading frame that contains the largest portion of the gene and does not mention the possibility of a shift in reading frame. This difference is illustrated by comparing the PGAP and RAST annotations of gene CA606_12630 (Figure 3). Taken together, the PGAP annotation is particularly useful because a frameshift qualifier search in Artemis allows the user to quickly locate possible sequencing errors that might have caused a change in reading frame; and while the RAST annotation does not offer this feature, it is useful when locating and verifying the reading frame for the larger gene fragment.

In addition to the frameshifted genes, PGAP identified 22 additional genes that harbored internal stop codons. However, RAST annotated 10 of the 22 genes with a start site that excluded the internal stop codon (Table S4). In all 10 cases, the start codons chosen by RAST were located at the beginning of the rise in third position GC content, which is typical of *Caulobacter* protein coding regions [6]. Database matches with the NA1000 reference genome also verified that the NA1000 start codons for nine of these genes matched those selected by the RAST annotation. In the other case, the start codon chosen by PGAP coincided with a small region of amino acid homology and then included the in-frame stop codon as shown by the example in Figure 4. The PGAP annotation would be useful if the genome sequence was likely to contain sequencing errors, but we were able to confirm that the nucleotide sequence was accurate in this region of the genome (see below). Thus, the RAST annotation probably represents the true start codon for this gene even though it is

different from that of NA1000; however, it is also possible that this gene may not be expressed in CB13. The remaining 12 genes with internal stop codons were absent in the RAST annotation and were not considered to be genes based on insignificant BLASTp database matches to the NA1000 genome and the Artemis GC frame plot peak (peaks aligning with the start of the gene). These results illustrate that the RAST annotation system locates accurate gene start sites more efficiently than the PGAP annotation system and that PGAP identifies gene fragments as potential genes. This identification of gene fragments could be useful for identifying the evolutionary history of particular genes in a comparison of several closely related genomes.

Since the PGAP annotation of the CB13 genome suggested that sequencing errors may be responsible for some of the annotated genes containing frameshifts, we decided to re-sequence the complete genome of CB13. Analysis of the re-sequenced CB13 genome demonstrated the exceptional accuracy of the PacBio sequencing system, since no SNPs were detected between the two versions of the genome sequence. However, when the homopolymer regions were compared, we found 93 single-base deletions in the original CB13 genome sequence and 16 single-base deletions in the second CB13 genome sequence. Since we have shown that the PacBio sequencing system generates sequencing errors by underestimating the number of bases present in a homopolymer region (Ely et al., manuscript in preparation), we manually added a single base to the original sequence in each of the areas denoted by the exported gap file. The corrected sequence was then submitted to GenBank and RAST to be re-annotated. The PGAP re-annotated CB13 genome contained 31 additional genes and 24 fewer pseudogenes for a total of 3,891 genes (Table 2). The RAST re-annotated CB13 genome had only 3,863 genes which is 18 fewer genes than the previous annotation. The increase in annotated genes by PGAP relative to the previous PGAP annotation can be explained by its reliance on predictive models based on open reading frame (ORF) searches, where small ORFs are often misinterpreted as actual genes. However, since none of the newly annotated genes included a sequence correction, this result also indicates that the PGAP annotation process is somewhat stochastic, and that new candidate genes may be identified each time a PGAP analysis is conducted. Our results show that neither annotation system is completely superior to the other method; however, these data support the findings that PGAP tends to over-annotate genomic data [9].

Of the 64 genes originally annotated by PGAP but not included in the RAST annotation, only 55 were included in the PGAP re-annotation, and the new RAST annotation included three of these genes. Similarly, of the original 16 genes annotated by RAST but not by PGAP, all except for one gene were included in the RAST re-annotation. Additionally, six of these 16 genes were included in the PGAP re-annotation. Despite these inconsistencies, these data support the notion that the PGAP annotation system is much more likely to detect novel genes and that the RAST annotation process is less stochastic.

The revised CB13 sequence corrected 31 of the 73 genes that the PGAP system annotated as frameshifted genes, which verified our hypothesis that some of the annotated frameshift genes were due to sequencing errors in homopolymer regions. However, the new PGAP annotation included 13 new genes that were annotated as frameshifted. Twelve of these were relatively small (~80 amino acids); did not have statistically significant (E-value $>10^{-5}$ and

query coverage <60%) matches in the GenBank database (Table S5); and had no definitive GC frame plot peak, so they were deleted from the annotation. In contrast, the 13th gene, CA606_07060, which encodes an ABC transporter protein, was originally annotated by the PGAP system without a frameshift but was re-annotated with a frameshift; however, this change in the annotation was not a direct result of base additions, since no bases were added in that region (Table S1). Surprisingly, the PGAP system correctly predicted the gene product for gene CA606_07060, but it failed to annotate the gene in the correct reading frame. In this case, differences in the CB13 sequence resulted in a start site for the CB13 gene that is different from that used for the corresponding gene in the NA1000 genome. However, a small nucleotide sequence (~ 50 bp) that corresponds to the 5'-end of the NA1000 gene is present in a different reading frame in the CB13 genome and was annotated as the start of the gene by PGAP. Collectively, these data highlight a pitfall within the PGAP annotation system and reinforce our idea that analyzing genomic data with more than one annotation system can provide a more accurate way to annotate nucleotide sequence data.

Of the 42 frameshifted genes that were not corrected by re-sequencing, 25 genes were relocated to a different reading frame, while 17 genes remained in the same reading frame even though a base modification was not introduced in the vicinity of the protein coding regions for 41 of these 42 genes. The remaining gene, CA606_20140, did receive a base addition within its protein coding region and remained annotated as a frameshifted gene in the PGAP annotation but was absent in the RAST annotation. BLASTp matches indicated significant matches for the beginning of this gene, which suggests that this gene possesses an actual frameshift that truncates the protein produced by this gene. Of the remaining 41 genes, 13 genes ranging from 200 bp to 3,368 bp were annotated in a reading frame that was completely inconsistent with the GC frame plot peak; however an ORF that corresponded to the GC frameplot peak was always directly above or below the annotated frameshifted gene and coded for all but the beginning of the amino acid sequence in the corresponding NA1000 gene. Thus, these 13 genes provide additional examples where PGAP incorrectly identified the reading frame of a gene due to sequence differences in the beginning of the gene. An additional seven genes corresponded with the GC frame plot peak more accurately when split into two separate genes. Interestingly, five of the seven genes encode non-hypothetical proteins with DNA replication functions such as endonuclease activity and DNA ligase activity. In each of these seven genes, only the beginning of the gene had significant protein matches to the NA1000 genome. Since we have confirmed the sequence of these genes, we concluded that they do contain frameshift mutations and that these genes are either not expressed or they produce nonfunctional peptides. Further, there are 12 additional endonuclease genes in the genome that are not annotated with a frameshift mutation, and there are two additional DNA ligase genes annotated without a frameshift mutation, which would circumvent the need for the frameshifted genes to produce functional proteins. The remaining five genes coded for proteins that were not essential for cellular function, but each gene did, however, have more than one copy present in the CB13 genome.

The remaining 21 frameshifted genes lacked a definitive GC frame plot peak and did not have a statistically significant match in the GenBank database, so they were removed from the annotation. Of the 21 genes, 10 genes matched only the beginning of the corresponding gene in the NA1000 genome; 8 genes matched the start sites of the corresponding gene in

the *C. crescentus* CB15 genome, but had been eliminated from the NA1000 genome annotation; and the remaining 3 genes had no matches to proteins in the database. These data are in agreement with our hypothesis that PGAP will identify more genes as frameshifted genes than the RAST system, because it tends to annotate initial gene fragments.

Collectively, these re-sequencing and re-annotation data illustrate that the PGAP system annotates partial genes more often in comparison to the RAST system and employs no corrective method to adjust the reading frame of the frameshifted gene or the start site of genes harboring an internal stop codon. But, in doing so, the PGAP annotation also accumulates a large number of pseudogenes, which may reflect evolutionary history but likely do not produce gene products. To highlight the genomic features derived from this comparative annotation analysis, a representative consensus chromosome image of CB13 was generated (Figure 5).

In summary, we have demonstrated the advantages of comparing annotations produced by two different annotation systems when analyzing genomic data. Our findings illustrate and reinforce that PacBio sequencing technologies generate very accurate full length bacterial genome sequences. However, re-sequencing the genome can be used to eliminate the residual sequencing errors present in some of the homopolymer regions. We have shown that the PGAP system is valuable for quickly locating frameshifted genes, which can help identify potential errors in the genome sequence that may inappropriately cause a gene to appear nonfunctional. In contrast, RAST is more efficient at annotating the correct gene start codons and produces fewer pseudogenes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This work was funded in part by National Institutes of Health grant R25GM076277 to BE.

References

- 1). Aziz RK, Bartels D, Best AA, et al. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. [PubMed: 18261238]
- 2). Christen B, Abeliuk E, Collier JM, et al. (2011) The essential genome of a bacterium. *Molecular systems biology* 7:528. [PubMed: 21878915]
- 3). da Silva CA, Lourenço RF, Mazzon RR, et al. (2016) Transcriptomic analysis of the stationary phase response regulator SpdR in *Caulobacter crescentus*. *BMC Microbiology* 16:66. [PubMed: 27072651]
- 4). Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5:e11147. [PubMed: 20593022]
- 5). Darling AE, Tritt A, Eisen JA, et al. (2011) Mauve assembly metrics. *Bioinformatics* 27:2756–2757. [PubMed: 21810901]
- 6). Ely B, Scott LE (2014) Correction of the *Caulobacter crescentus* NA1000 genome annotation. *PLoS one* 9:e91668. [PubMed: 24621776]

- 7). Kislyuk AO, Katz LS, Agrawal S, et al. (2010) A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* 26:1819–1826. [PubMed: 20519285]
- 8). Marks ME, Castro-Rojas CM, Teiling C, et al. (2010) The genetic basis of laboratory adaptation in *Caulobacter crescentus*. *J Bacteriol* 192:3678–3688. [PubMed: 20472802]
- 9). Nielsen P, Krogh A (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* 21:4322–4329. [PubMed: 16249266]
- 10). Overbeek R, Olson R, Pusch GD, et al. (2013) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 42:D206–D214. [PubMed: 24293654]
- 11). Pruitt KD, Tatusova T, Brown GR, et al. (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 40:D130–D135. [PubMed: 22121212]
- 12). Rutherford K, Parkhill J, Crook J, et al. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 10:944–945.
- 13). Schrader JM, Li GW, Childers WS, et al. (2016). Dynamic translation regulation in *Caulobacter* cell cycle control. *Proceedings of the National Academy of Sciences* 113:E6859–E6867.
- 14). Scott D, Ely B (2015) Comparison of genome sequencing technology and assembly methods for the analysis of a GC-rich bacterial genome. *Current Microbiol* 70:338–344. [PubMed: 25377284]
- 15). Shin SC, Ahndo H, Kim SJ, et al. (2013) Advantages of single-molecule real-time sequencing in high-GC content genomes. *PLoS One* 8:e68824. [PubMed: 23894349]
- 16). Tatusova T, DiCuccio M, Badretdin A, et al. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* 44:6614–6624. [PubMed: 27342282]

Figure 1.

A graphical representation of the annotated gene discrepancies between the genes identified in the PGAP annotation (outer circle) and those identified in the RAST annotation (inner circle) of CB13. As shown, the majority of the annotation differences are among genes that code for hypothetical proteins; however, a portion are related to cellular regulation [3]. These differences may be explained by the usage of GeneMarkS+ by PGAP, whereas RAST primarily relies on sequence similarity homology.

Figure 2.

A Mauve alignment snapshot depicting a representative annotation inconsistency between the PGAP annotation (top) and the RAST annotation (bottom) of the CB13 genome. While the PGAP annotation includes gene CA606_07250, the RAST annotation excluded this gene, which is represented by the missing gene enclosed by the black box.

PGAP

RAST

Figure 3.

An Artemis snapshot image showing the frameshifted gene (CA606_12630) annotated by PGAP using only the top reading frame where the gene starts and does not indicate that the reading frame shifts to the middle reading frame after the addition of an extra base. In contrast, the equivalent gene (CDS blocked in bold) annotated by RAST uses the reading frame that corresponds to the end of the gene after the additional base, which changes the reading frame. Note that the GC frameplot graph indicates peaks in both reading frames.

Figure 4.

An Artemis snapshot alignment illustrating the internal stop (line in the middle of the gene enclosed by the box) present in the PGAP annotation of gene CA606_12670. The equivalent gene in the RAST annotation uses a start codon downstream of the internal stop, which is the correct annotation based on a BLASTp homology analysis and by the position of the corresponding third position peak generated by the GC frame plot feature of Artemis. Note that this annotation is in closer agreement with the GC frameplot peak in comparison to the start position of the PGAP annotation.

Figure 5.

A representation of the modified annotation of the CB13 genome combining information from both the RAST and the PGAP annotations. The outermost concentric rings represent the genes present in the forward and complementary strands, respectively. The third ring shows the positions of the pseudogenes from the PGAP annotation. The innermost concentric ring represents the GC content fluctuations, with lighter gray representing regions of lower GC content. The solid gray circle merely separates the two sets of information.

Table 1.

Original CB13 sequence annotations.

	PGAP	RAST
Unique genes	64	16
Total annotated frameshifted genes	73	1
Genes annotated with an internal stop codon	22	0
Genes with an internal stop requiring a start codon modification	10	0
Genes with an internal stop that are probably not genes (deleted genes)	12	0

Table 2.

CB13 genome feature differences between PGAP and RAST.

Genomic Feature	PGAP (original)	PGAP (re-sequenced)
Number of bases	4,143,958	4,144,051
Number of genes	3,836	3,891
tRNA genes	51	51
Pseudogenes	135	111
GC%	67.13	67.13
	RAST (original)	RAST (re-sequenced)
Number of bases	4,143,958	4,144,051
Number of genes	3,881	3,863
tRNA genes	51	51
Pseudogenes	0	0
GC%	67.13	67.13