

Abstract

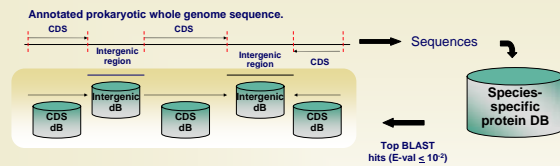
Manual annotation is essential for the curation of whole-genome sequences. Several analytical challenges make the follow-up curation of automatically-annotated genomes a slow and labor-intensive process. The first is the identification of polymorphisms resulting in frameshifts. In bacteria, genes with disrupted reading frames due to frameshifts are often mis-annotated as two genes (a "split gene"). Unfortunately, in the absence of additional experimental evidence, it is difficult to determine whether a split gene is the result of sequencing error or a legitimate polymorphism. Another challenge is the identification of genes missed by automated annotation. Developing an efficient analysis system, which can associate the sequence variants with pathogenicity, immunogenicity and other organism-specific properties of bacterial pathogens, could pose as an additional challenge. Herein, we present a context-based, polymorphism oriented analysis methodology (Gen-Var) for genome annotation improvement and comparative analysis. Gen-Var, developed to address the annotation challenges, is based on GeneWise¹ and is designed to automatically identify sequence variants and missed gene assignments, and to comparatively analyze these genome features in the context of closely related organisms. Results from the analysis of four sequenced *Brucella* genomes indicate that the method can significantly improve genome annotation. For example, the Gen-Var analysis detected about 339 previously unidentified split genes in *B. melitensis* 16M and revealed that about 179 (53%) of them have likely resulted from sequencing errors. The analysis also revealed many genes missed by earlier annotation efforts, species-specific gene disruptions and modifications and their potential roles in the virulence of *Brucella*.

Motivation

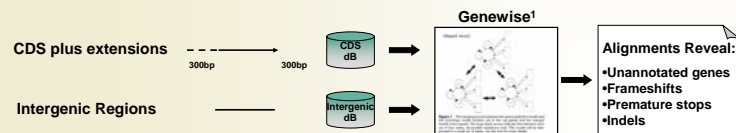
- Exponential increase in the number of prokaryotic whole genome sequences.
- Generating high-quality data from genomic sequences is key.
- Annotation "Gold Standard" = Manual annotation, but too time consuming
- Common annotation problems addressed by this research:
 - Standardize annotations across related genomes.
 - Resolve conflicts in annotations.
- Address these challenges
 - Provide consistent and standardized annotations across related genomes.
 - Identify frame-shifts and premature stop codons (real? sequencing errors?).
 - Identify and map insertion-deletion events (indels) in the annotations of whole genome sequences.

Approach

Step 1: Using BLAST, create a mini-database for each CDS and intergenic region.



Step 2: Use GeneWise¹ to align each sequence with the sequences in the mini-database



Application Example: *Brucella*

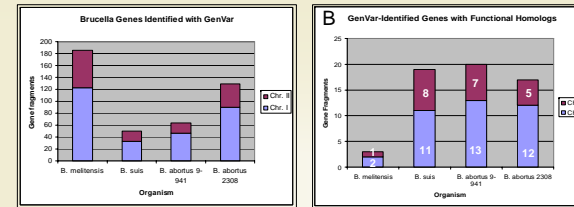
Brucella is an alpha-proteobacterial genus containing several pathogenic species relevant for human and animal health. We analyzed four *Brucella* genomes:

- *B. melitensis* 16M
- *B. suis* 1330
- *B. abortus* 2308
- *B. abortus* 9-941

The species specific database for this analysis included sequences from the following:

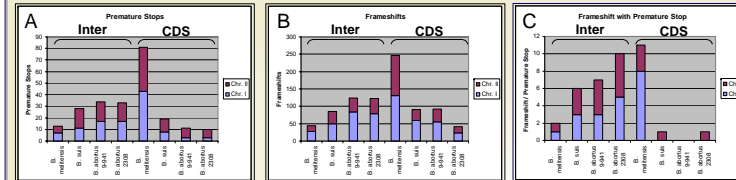
- The four *Brucella* genomes
- *E. coli* K12 (very well annotated, gamma-proteobacterial genome)
- Uniprot (wider diversity)
- *Agrobacterium tumefaciens* C58 (alpha-proteobacterium)
- *Mesorhizobium loti* MAFF303099 (alpha-proteobacterium)

Missed Gene Identification



Missed gene assignments revealed in the intergenic DNA regions from the four *Brucella* genomes. The bars illustrate the total number of novel gene identifications (Panel A) and the number of these novel gene calls that are longer than 100 AA and have good DB hits with genes with well-defined biological functions in other organisms (Panel B). Each chromosome was analyzed individually. The bars show the breakdown between the contributions from Chromosome I (blue) and Chromosome II (maroon) for each genome, and the numbers within each bar show the number of genes identified in each chromosome.

Split Genes



GenVar identified split genes in all four *Brucella* species. The genes identified were categorized as to whether they contained a premature stop codon (Panel A), a frameshift (Panel B), or a frameshift that resulted in a premature stop codon (Panel C). This analysis was done separately for genes identified within regions previously annotated as intergenic (left side of each panel) and within protein coding sequences (right side of each panel). Interestingly, there are 8 corresponding split genes that occur in pairs of *Brucella* species (*B. melitensis* and *B. abortus*, or *B. suis* and *B. abortus*). There are also 56 split genes that occur in both *B. abortus* species. It is unlikely that a gene with a split at the same location in multiple genomes is the result of a sequencing error.

Indels



Over 140 indels (insertions or deletions of multiple residues as shown by a multiple alignment) were identified by GenVar in *B. melitensis* 16M, with comparable numbers in the other species. Two examples are shown here. The sensor histidine kinase in *B. abortus* 2308 contains an insertion of 4 residues at position 422 relative to its orthologs in the other *Brucellae*. *B. suis* also contains a deletion of 4 residues relative to the orthologs in the other *Brucellae*. The VirB10 protein of *B. melitensis* 16M contains an 8 residue deletion at position following amino acid 143, relative to *B. suis* and *B. abortus* 2308 and amino acid 147 relative to *B. abortus* 9-941.

Conclusion

GenVar has been developed to address the need for an efficient analysis system to improve genome annotation. It automates the identifies the following genomic features: **Missed genes, split genes, indels**. GenVar also automates the comparison of sequence variants that may be scientifically interesting for pathogenicity, immunogenicity and other organism-specific properties of bacterial pathogens. In addition to *Brucella*, GenVar is being used to analyze the genomes of *Rickettsia* and *Coxiella* species, which are within the scope of the PathoSystems Research Integration Center [PATRIC², <http://patric.vbi.vt.edu>]. Results will be made publicly available within a few months.

References

- Birney, E., M. Clamp and R. Durbin (2004). "GeneWise and Genomewise." *Genome Res* 14(5): 988-95.
- Snyder, E. E. et al (2007). The VBI PathoSystems Resource Integration Center (PATRIC). *Nucleic Acids Research*, in press.

Acknowledgment

This project is funded by NIAID / NIH contract HHSN26620040035C to Bruno Sobral. Thanks to June Mullins for poster graphic design and layout.