

TITLE: **Analysis and Curation of Overlapping Gene Models**

PAGE: 1 of 2

SOP #: PA04

REVISION LEVEL: .2

EFFECTIVE DATE: July 2006

AUTHOR:  
Lauren Brinkac

PRIMARY REVIEWER:  
Bill Nelson

## 1 OVERVIEW

It is generally accepted that genes in prokaryotes do not overlap significantly. Thus when two gene models collide, it is assumed that either one or both of the models is incorrect. Overlapping gene models are resolved either by choosing an alternate initiation codon for one or both models, or determining on model to be a false positive and deleting it.

### 1.1 Scope

This analysis is performed on all gene models that have overlapping regions of 30 nucleotides or more.

### 1.2 Related Documents

SOP PA01 - Gene Prediction

SOP PA02 - Homology Searches

SOP PA03 - Start Site Curation

### 1.3 Revision History

| Author      | Date         | Change                           |
|-------------|--------------|----------------------------------|
| Bill Nelson | 14 July 2006 | Basic edits and addition of SOP# |

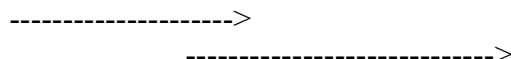
## 2 REQUIREMENTS

A genome sequence with gene model predictions and homology searches run.

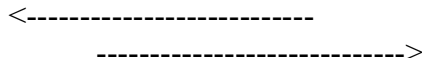
## 3 PROCEDURE

### 3.1 Type of gene model overlap

3.1.1 The predicted genes are in the same orientation with the 3' end of the first overlapping the 5' end of the second.



3.1.2 The predicted genes are in the reverse orientation with their 5' ends overlapping.



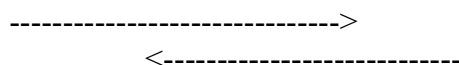
TITLE: **Analysis and Curation of Overlapping Gene Models**

SOP #: PA04

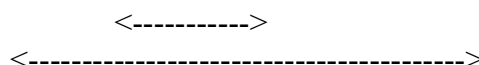
REVISION LEVEL: .2

PAGE: 2 of 2

3.1.3 The predicted genes are in the reverse orientation with their 3' ends overlapping.



3.1.4 One predicted gene is entirely contained within the other gene.



### **3.2 Criteria for overlap editing**

Overlapping gene models in categories 3.1.1 and 3.1.2, and sometimes category 3.1.4, can often be resolved by standard start site analysis. If start site analysis fails to resolve the overlap, one of the overlapping models is considered for deletion. In general, gene models for which there is an assigned identity and/or sequence similarity to some other protein or domain are retained, while those models lacking such evidence are regarded as hypothetical and are candidates for deletion. The following factors are considered in this decision:

#### 3.2.1 Predicted gene length

In general, the longer the open reading frame (ORF), the more likely it contains a real gene. Glimmer predicts genes as small as 90 nt, which are frequently annotated as “hypothetical proteins” (see Gene Naming Conventions). Overlaps are often resolved by the deletion of these smaller (90-150 nt) hypothetical genes. As gene size increases more caution is used. One caveat to this rule is the phenomenon of 'shadow ORFs', ORFs on the opposite strand that inhabit the same stretch of sequence. These are common in GC-rich genomes, and can be quite long, but usually have no homology to known proteins. Gene predictions within these ‘shadow ORFs’ are deleted.

#### 3.2.2 Genome context

If the majority of gene models in a region appear to be in the same orientation and have similar functional roles, then an overlapping hypothetical protein in the opposite orientation is deleted.

TITLE: **Analysis and Curation of Overlapping Gene Models**

SOP #: PA04

REVISION LEVEL: .2

PAGE: 3 of 2

### 3.2.3 Confidence of identity

Although most overlaps involve hypothetical proteins, some gene models get functional identifications based on hits to low complexity sequence or to inaccurate gene calls. In these cases, if the identification of a gene model is based on erroneous evidence, the model is deleted. In other cases, there may be hypothetical proteins that are members of paralogous families, or have PROSTE motifs. The presence of this additional information may provide enough evidence for that gene model to be retained.

### 3.2.4 Gene composition

Gene models that have few possible start codons, all of which are the low frequency GTG and TTG codons are considered weak gene calls. In GC-rich organisms, 3<sup>rd</sup> position GC skew is a good indicator of coding sequence. Gene models that have predicted translations with low complexity sequence or unusual composition are also viewed negatively.

## 4 DATA MANAGEMENT

### 4.1 *Quality Control*

The final gene list is re-evaluated for overlaps. Any remaining overlaps of more than 30 nucleotides are reviewed for accurate analysis.