

## **PATRIC Standard Operating Procedures for Manual Nucleic Acid-Level Curation of Bacterial and Viral Genomes**

---

**Version:** 1.0  
**Date:** July 18, 2006  
**Supercedes version:** N/A

### **Abstract**

Manual genome curation is the process whereby members of the curation team analyze any preexisting genome annotations in the context of the data generated by the Genomic Sequence Analysis Pipeline (GSAP). Automated evaluation of this data prioritizes curation for those genes where conflicting evidence is present and human judgment is required. These standard operating procedures (SOP) provide guidance on how to interpret the data presented to the curator.

### **1. Rationale**

The genomic (DNA level) curation of bacterial and viral genomes is a very time consuming process. The majority of the cases are straight forward and do not need any manual modifications. The aim of automated curation of bacterial genomes is to automatically annotate these straight forward examples, without sacrificing the quality attained by manual, expert curation (the gold standard). The genes that are not straight forward to annotate are hand curated.

### **2. Methods**

The starting point of our annotation is currently either a GenBank entry or RefSeq from NCBI, which is analyzed by the Genome Sequence Annotation Pipeline (GSAP). Gene prediction programs, Glimmer and GeneMark, are used to substantiate the existing annotations and/or to identify putative new genes. We use some other programs to predict the correct start sites (like, Ribosomal Binding Sites finder (RBS finder), TICO and Start site consensus (SSC)).

### **3. Manual Annotation of Bacterial Genomes**

Genes assigned a curation status of interim in the automated curation pipeline are manually annotated using evidence of BLAST, RBS finder, and start site-consensus. Other features like RNAs, rRNAs, non-coding features, pseudogenes, and frameshifts are analyzed here.

#### **3.1 Start site correction**

The evaluation will result in either accepting the existing annotation or modifying (extending or shortening) the start site of the original annotation based on following criteria. A full-length (BLAST) alignment with orthologs, especially experimentally verified orthologs or those from reference genomes (in case of associate genome curation), will be given the highest weight to resolve the start site difference. The start sites predicted from TICO and RBS finder will be given a lower level of priority than BLAST.

The start-site of the existing annotation agrees with one of the predictions Glimmer or GeneMark and there is a BLAST hit (to the existing annotation) to an ortholog with same start site, then start site is not corrected. If the existing annotation disagrees with both Glimmer and GeneMark, the BLAST output is analyzed. If there is a BLAST hit to an ortholog and its start matches with the existing annotation, the start site is not corrected.

If the existing annotation disagrees with both Glimmer and GeneMark, the BLAST output is analyzed. If there is a BLAST hit to an ortholog and its start matches with either Glimmer or GeneMark but not with the existing annotation, the start site is corrected to that of BLAST hit. If the ortholog identified by BLAST does not match with Glimmer, GeneMark and the existing annotation, the start site is corrected using the TICO suggested correction. In this case the longest TICO-corrected sequence of Glimmer, GeneMark or RefSeq is accepted.

In the absence of BLAST hits the longest of RefSeq, Glimmer or GeneMark after applying TICO correction is applied.

### **3.2 Gene Deletion**

Whenever two existing annotation are present in the same region and if there is a significant overlap between them the shorter one is deleted. The longer one is evaluated by the criteria outlined in above sections.

In the absence of an existing annotation, if there is a GeneMark or Glimmer prediction the longer one is considered and if there is no BLAST hit, the predicted sequence is analyzed by InterProScan to identify any significant domains. If there are no InterProScan domains the prediction is deleted.

### **3.3 New Genes**

In the absence of an existing annotation, the longer of a GeneMark or Glimmer prediction is considered. If there is a BLAST hit, the prediction is accepted as a new gene after applying TICO correction.

In the absence of an existing annotation, the longer of a GeneMark or Glimmer prediction is considered and if there is no BLAST hit, the predicted sequence is analyzed by InterProScan to identify any significant domains. If domains are identified by InterProScan it is accepted as a new gene.

### **3.4 Pseudo-gene**

The pseudogene will be annotated in accordance with the literature citations.

### **3.5 Entering Comments**

RBS finder and SSC are considered as supporting evidence and added in the comments section

### **3.6 Premature Stop Codons**

The following observation in the process of manual curation may indicate premature stop codon in the genes:

- When a BLAST alignment with a gene reveals stop codons within the open.

### **3.7 Frameshift Detection**

The following observations in the process of manual curation may indicate frameshifts in genes:

- When a gene aligns with two or more BLAST hits in different reading frames on the same genome strand, a possible frameshift or sequencing error may be indicated.
- When a small portion of the gene is missing at either end, a detailed comparison between their aligned genes and three open reading frames corresponding to the missed portions may reveal possible frame shifts.

## **4.0 Manual Annotation of Viral Genomes**

Manual annotation of viral genomes is less straightforward than annotation of bacterial genomes. Diversity in viral sequences, the state of biological understanding, and the availability of published literature for each viral class necessitates that slightly different approaches be taken for each viral class. The below SOP provides the general approach to viral curation.

### **4.1 Reference Genome:**

In each viral class, one or more genomes are selected as reference genomes in consultation with organism experts, based on their biological importance and ability to represent a category (such as genus) of virus. The remaining associated genomes are annotated based on reference genome annotations.

#### **4.1.1 Coding features:**

For the genomes with GenBank/RefSeq annotations:

The coding features are checked against GeneMark predictions and BLAST homologies and if all agree, the GenBank/RefSeq features are selected. And if the GeneMark predictions or BLAST homologies do not agree with the GenBank/RefSeq feature, the GenBank feature is taken as the default feature.

For the genomes that are not annotated in GenBank:

A program, GATU (Genome Annotation Transfer Utility) is used to transfer features from the annotated genomes. This program performs BLASTp searches and pairwise alignment of features in the annotated genome (referred to as reference genome in this program) against the genome to be annotated (referred to as associated genome in this program). The features from this program are selected based on the similarity percentage (i.e., the similarity percentage should be 85% or more or the similarity threshold can be determined based on the evolutionary closeness of the genomes being annotated). Phylogenetic relationship between the genomes is used to determine which genomes should be used as the reference genome to annotate the un-annotated genomes.

#### **4.1.2 Non-coding Features:**

The non-coding features are annotated from the literature. The set of non-coding features described will depend upon the virus; its biology, and what is known about its non-coding features. The features that have been described for PATRIC viruses include the following: 5' UTR, 3' UTR, consensus mRNA start sites, leader RNA, alternate in-frame start codons, polypyrimidine tracts, IRES, and poly A tracts.

## **4.2 Associated Viral Genome**

### **4.2.1 Coding Features**

For the genomes with GenBank/RefSeq annotations, the coding features are checked against the GeneMark predictions and BLAST homologies and if all agree, the GenBank/RefSeq features are selected. And if the GeneMark predictions or BLAST homologies do not agree with the GenBank/RefSeq feature, the GenBank/RefSeq feature is taken as the default feature.

For the genomes that are not annotated in GenBank, the GATU program is used to transfer features from the reference genomes to the un-annotated genomes (as mentioned in section 4.1.1).

#### **4.2.2 Non-coding Features**

Multiple sequence alignment of the non-coding regions is used to transfer non-coding features from the reference genome to the associated genomes.

#### **5.0 Provenance**

The Curation Tool and PATRIC database also support mechanisms to communicate inference. For example, if a curated gene feature is based on a particular prediction program and a specific literature reference, it is possible to encode that fact along with the feature. This allows a user to determine on what basis a particular annotation was made. There are several reasons for doing this. If later work reveals a premise to be faulty, one should be able to identify the information on which the original curation was made. In addition, as new data is added, this provides a mechanism to identify new information that should be considered in the context of existing curated features.