

# PATRIC Standard Operating Procedures for Automated DNA-Level Curation (ADC)

---

Author: Eric E. Snyder  
Version: 1.0, October 2, 2006 10:09 AM  
Supersedes version: N/A

---

## 1. Purpose

This SOP describes the Automated DNA-Level Curation system (ADC), a preprocessor for microbial genome annotation, used to process output from the Genomic Sequence Annotation Pipeline (GSAP) prior to manual curation. The goal of this document is to familiarize personnel with the configuration, execution and output of the ADC system.

## 2. Application

Although this SOP is narrowly focused on its application to the NIAID/BRC PATRIC project, it will be helpful to other groups interested in using the Curation Infrastructure developed by the Genomics Domain Area Team at VBI. When used in concert with GSAP, the system is flexible and can be readily adapted to analyze the genomes of other organisms, or other types of sequence data.

## 3. Participants

This SOP is typically executed by a software developer working with a bioinformatician following the execution of GSAP.

## 4. Inputs

The ADC application requires as input the name of the CI database containing the genomic data of interest and the name of that genome.

## 5. Entry Criteria

To execute the processes described herein, access to the ADC application and the CI database containing a complete genome record including status as a "reference" or "associated" genome and legacy annotation from the original RefSeq or GenBank record. The genome must have been processed by GSAP, which implies that specific applications such as BLASTX (vs. UniProt) and various gene prediction programs have been executed.

## 6. Automated Gene Curation in Bacteria

GSAP generates all the evidence require by the curators to make their final gene calls. Unfortunately, this is a tedious manual process which requires attention to detail and is subject to significant inter-curator variability. To speed up this process and reduce the amount of manual intervention and its consequences, an automated curation system has been developed. The goal of this system is to identify coding regions and classify them into three groups based on GSAP features and primary annotation. The groups indicate "curation status", which is a reflection of the confidence in each CDS assignment and the disposition of the CDS with respect to further curation. Curation status values and their interpretation are described in Table 1.

**Table 1: Curation Status**

Status	Meaning
Final	The coordinates of the feature are accepted as being correct. The feature is withdrawn from the curators' work list.
Pending	The feature is probably correct but has limited supporting evidence and has been flagged for review by the curation staff.
Interim	The feature has only marginal supporting evidence and has been flagged for a thorough review.

Current results indicate that 50% to 75% of coding regions can be promoted to “pending” using the automated curation system (described below).

## 6.1 Reference Genomes

Using the output of GSAP, CDS features are clustered on the forward and reverse strands based on their 3'-ends (stop codons). CDS coordinates predicted by Glimmer, GeneMark and TICO-corrected<sup>1</sup> Glimmer and GeneMark are considered along with the primary GenBank or RefSeq annotation and their TICO corrections. If at least three data sources agree and one is the primary annotation, the coordinates and data associated with the primary annotation are promoted to “final”. Other situations are flagged as “pending” and queued for manual follow-up as described in Table 2.

**Table 2: GSAP Automated Curation Process for Reference Genomes** This process examines the features produced by GSAP and evaluates the prediction most likely to represent the actual coding region (the gene’s CDS). Based on the level of supporting evidence, a “curated CDS” feature may be created and assigned a curation status, as described in **Table 1**. Here, the numbers below the program name indicate a particular set of coordinates so that they may be compared to others on the same horizontal line. Thus, in case 1.0, the coordinates from GenBank (or RefSeq, depending on the annotation’s origin), Glimmer, GeneMark and TICO (*i.e.* the CDS reflecting the TICO corrections) all agree, while the presence of BLAST alignments was not considered. In cases 6.0 – 6.3, +1 and -1 indicate positive and negative strand predictions, respectively. Square brackets are used to indicate a logical “or”; for example, in case 3.0, if the TICO prediction is the same as 1, 2 or 3, then the matching prediction is used to create the curated feature. In other words, the TICO prediction is “promoted” and assigned a status of “pending”. In situations where BLASTX information is used, *each* prospective CDS is aligned to the non-redundant protein database. BLASTX results are considered positive when an alignment exceeds the threshold of  $E < 10^{-5}$ . This is a liberal threshold; the actual alignments are inspected on manual follow-up.

Curation Status	Case	GenBank	Glimmer	GeneMark	TICO	BLASTX	Promote the following:
Final	1.0	1	1	1	1		GenBank
	1.1	1	1	1	2		GenBank
	2.0	1	1	2	1		GenBank
	2.1	1	2	1	1		GenBank
Pending	2.2	1	2	1	2		GenBank
	2.3	1	1	2	2		GenBank
Interim	3.0	1	2	3	[1,2,3]		TICO
	3.1	1	2	3	4		longest
	4.0	-	1	-		+	Glimmer
	4.1	-	-	1		+	GeneMark
	4.2	-	1	-		-	Glimmer
	4.3	-	-	1		-	GeneMark
	5.0	-	1	1	1	-	Glimmer
	5.1	-	1	1	2	-	TICO
	5.2	-	1	2	[1,2]	-	TICO
	5.3	-	1	2	3	-	longest
	6.0	-	+1	-1	+1	+	Glimmer
	6.1	-	-1	+1	-1	+	GeneMark
	6.2	-	+1	-1	[0,-]	+	Glimmer
	6.3	-	-1	+1	[0,-]	+	GeneMark
	7.0	-	1	-		-	Glimmer
7.1	-	-	1		-	GeneMark	

<sup>1</sup> The translational-start site correction program, TICO, takes a genomic DNA sequence plus one or more CDS features (*e.g.* gene predictions, primary CDS annotations, *etc.*) as input and generates a list of corresponding “corrected CDSs”, which take into account the authors improved weight matrix based on triplet frequencies in the vicinity of known start sites.

## 6.2 Associated Genomes

A second decision tree has been implemented for bacterial AGs. The Reference Protein List is aligned one protein at a time to the AG using TBLASTN to identify orthologous genes. The coordinates of the alignment on the AG are processed further to identify valid start and stop codons. The 5'-end is identified using a log-likelihood position-weight matrix for the start codon consensus for the organism in question. Starting at the 5'-end, the weight matrix is scanned in 3 bp increments up-stream and down-stream, in an alternating fashion, for a maximum of 15 bp downstream and 99 bp upstream, until a valid site (matrix score > 0) is found. A similar scanning approach is used to identify the stop codon associated with the alignment. Coding sequences identified in this manner are identified as "RG-BLAST" features. If no termini meeting these requirements can be found, the alignment is kept but no RG-BLAST feature is created.

Following the RPL alignment step, CDS features from RG-BLAST and GenBank/RefSeq (hereafter, "RefSeq"), Glimmer, GeneMark plus their TICO-corrected counterparts are clustered as described in section 6.1. If an RG-BLAST feature is present in a cluster and the start site of the RefSeq annotation is within 99 bp, the RG-BLAST feature is given a status of "final" and is not reviewed manually. If the start site differs by more than 99 bp, the RG-BLAST feature is given a status of "pending", meaning that it is targeted for manual follow up. When RefSeq is available in the absence of RG-BLAST, the longest CDS is promoted to "final" if the difference between them is less than or equal to 99 bp. If it is more than 99 bp, the RefSeq feature is promoted to "pending". In the absence of RG-BLAST and RefSeq, the longest CDS is taken and promoted to "pending" status. This decision tree is illustrated in Table 6.

**Table 3: Decision Tree 2**

Cases:	1	2	3	4	5	6
RG-BLAST	+	+	+	-	-	-
RefSeq	+	+	-	+	+	-
Glimmer	+/-	+/-	+/-	+/-	+/-	+/-
GeneMark	+/-	+/-	+/-	+/-	+/-	+/-
RefSeq + TICO	+/-	+/-	+/-	+/-	+/-	+/-
Glimmer + TICO	+/-	+/-	+/-	+/-	+/-	+/-
GeneMark + TICO	+/-	+/-	+/-	+/-	+/-	+/-
Length Difference:	< 99 bp	> 99 bp	N/A	< 99 bp	> 99 bp	N/A
Choice:	RG-BLAST	RefSeq	RG-BLAST	Longest	Longest	Longest
Status:	Final	Pending	Final	Final	Pending	Pending

## 7. Exit Criteria

Following ADC execution, the database is scanned for curated features with statuses that do not conform to allowed values ("final", "pending", "interim" for reference genomes, "final" and "pending" for associated genomes). Similarly, features are checked to ensure that "case" values are recorded and correspond to the defined outcomes. In the event of anomalous behavior, the events are reviewed by the ADC developer for debugging and the run repeated, if necessary.

## 8. Performance Measures

After each ADC run, the standard battery of feature integrity checks is run. For example, all predicted CDSs should have length%3 = 0. Histograms of CDS feature length are plotted to identify outliers for manual inspection. In addition, queries are run to compare predicted CDSs to legacy annotation (GenBank or RefSeq) and to each other. This allows curators to assess the quality of previous annotation with respect to GSAP and generate a baseline against which to compare the results of manual annotation. The performance of individual tools can also be assessed to aid in their interpretation.