# Prokaryotic Automatic Annotation Pipeline SOP
# J Craig Venter Institute

**Author**: Ramana Madupu
**Version**: 1.01
**Effective Date**: 4/1/2009

## 1. Abstract

This SOP describes JCVI's Prokaryotic Genome Annotation Pipeline that annotates complete and draft genome sequences. JCVI's annotation pipeline is designed to identify an extensive collection of genome features, protein-coding regions, RNAs, regulatory features, repeat regions, and mobile genetic elements. Following identification, the predicted gene models are further refined using diverse evidence data types. Next, the functional annotation module assigns gene product names, where each evidence data type is ranked using precedence-based rules that favor more trusted annotation evidence to assign preliminary automated annotation to every gene model.

## 2. Introduction

The purpose of JCVI's annotation methods is to produce consistent, high quality automated annotation of complete and draft prokaryotic genomes. Among the challenges to consistent and accurate functional annotation of genes is the propagation of un-trusted and often inaccurate annotations via sequence homology to proteins in the public domain. One area of development of particular value is ongoing engineering improvements that will enable JCVI's pipeline to effectively integrate the growing body of high quality evidence and leverage experimental characterization of proteins into our annotation systems. JCVI is working towards scaling the capacity of the software to more effectively handle the deluge of prokaryotic genomes being sequenced with new sequencing technologies.

## 3. Requirements

### 3.1 Data requirements

The pipeline can process whole Genome sequences (single fasta file) or draft assemblies (multi fasta files). A pseudomolecule is generated when an unclosed molecule or draft sequence is entered into the JCVI annotation pipeline. The draft assemblies are stitched together into a pseudomolecule using a linker sequence, "NNNNNCACACACTTAATTAATTAAGTGTGTGNNNNN", which places start and stop codons in all 6 reading frames. Pseudomolecules generate better gene calls on contigs, especially when predicting partial genes at the ends of contigs.

### 3.2 Software requirements

Several open source algorithms, programs, inhouse scripts, data libraries are required and referenced in the procedures section

### 3.1 Compute requirements

Blast Extend Repraze (BER) and searches against Hidden Markov model libraries are performed on a High Throughput Computing (HTC) grid.

## 4. Procedure

**4.1 RNA Feature Identification:** A search for tRNA genes is performed using tRNAScan-SE (1). Ribosomal RNA genes, non-coding RNA genes and cis-regulatory RNA features such as riboswitches, are identified directly from BLAST search results or from matches to prokaryote-relevant subset of Rfam, a database of non-coding RNA families (http://www.sanger.ac.uk/Software/Rfam/). Additionally, tRNAs are confirmed and tmRNAs are identified using the ARAGORN program (http://130.235.46.10/ARAGORN1.1/HTML/). This step delineates regions of the genome for exclusion from protein-coding gene model insertion by ab initio gene finders.

**4.2 Protein-Coding Region Identification:** Protein-coding gene models are generated in a two-step process of *ab initio* prediction followed by evidence-based refinement. For complete or draft prokaryotic genomes the Glimmer3 (*2*) algorithm is used to obtain an initial set of gene predictions. The resulting gene set is supplemented through JCVI's ValetPep, a tool that detects biologically relevant, short gene models, based on sequence homology to highly trusted protein family models, commonly missed by *ab initio* gene prediction programs. Regions of the genome where gene features are either absent, or gene models lack biological evidence, are searched by BLASTX (*41*) against a database of non-redundant and non-identical proteins. Homology to known proteins is used to refine the start sites (see BER against PANDA, below). The JCVI tool AutoGeneCurate uses a combination of homology-based evidence types to extend gene models where needed and resolve conflicts between gene models by shortening or removing genes as necessary. Homology evidence for other potentially overlapping genes sets a no-further-upstream limit, while HMM or phylogenetically diverse BLAST (3)

homology evidence within the called gene sets a no-further-downstream limit. These limits converge to report that there is only one possible start site.

**4.3 Repeat Regions and Mobile Elements:** JCVI currently uses the ISfinder database (4) to identify insertion sequence (IS) elements, the PhageFinder algorithm (5**)** to identify prophage regions and PILER-CR (6) and CRISPRFinder (7) to find CRISPR repeats.

### 4.4 Homology Searches

#### 4.4.1 Protein and Nucleotide Data Archive (PANDA)

PANDA is JCVI's internal repository of non-redundant and non-identical protein and nucleotide data built periodically from public databases that include the latest protein sequences (e.g. GenBank (http://www.ncbi.nlm.nih.gov),PDB (http://www.rcsb.org/pdb/Welcome.do), UniProt (http://www.pir2.uniprot.org/), and the Comprehensive Microbial Resource database (http://www.tigr.org/CMR)).

**4.4.2 Characterized Protein Database (CHAR):** CHAR is a collection of experimentally characterized proteins and is ranked as the most trusted source for automated annotation. The CHAR database stores information on characterized proteins derived from the literature with standardized gene nomenclature linked through unique accessions to corresponding sequence entries in public databases. Each entry in CHAR is assigned Gene Ontology (GO) (8) function and process terms, Gene Ontology (GO) evidence codes, functional protein name, Enzyme Commission (EC) number (9) and Transport Classification (TC) numbers (50), gene symbol, and alternate names.

**4.4.3 HMM Searches Against Trusted Protein Families:** All proteins are searched against an HMM database of protein family models called TIGRFAMs (10) and PFAMs (11) using HMMER (12). TIGRFAM's focus is to build HMMs at the equivalog level; such HMMs model one specific function in the cell. Proteins that score well to these models can be assumed to share the same function that the HMM models. In addition to the TIGRFAM models, our HMM database also contains the Pfam set of models.

**4.4.4 BER Against PANDA:** A pair wise protein search program written at JCVI called BLAST Extend Repraze (BER) (http://ber.sourceforge.net) is used to search against PANDA. The same BER search provides homology-based evidence for functional annotation, where evidence is lacking from CHAR and trusted protein families, and is also an essential element of the refinement of gene structure. In BER all predicted proteins are extended by 300 nucleotides at both ends and a modified Smith-Waterman alignment (13) is

built between the extended version of the query sequence and the significant BLAST matches. BER searches in other frames and past stop codons for regions of similarity between two proteins. Therefore, if a sequencing error or natural mutation has split one gene into two (by a frameshift or in-frame stop codon) BER creates an alignment across those two fragments. BER is an exceptional method of detecting disrupted CDSs in genome proteins, and refining start site predictions.

**4.4.5 Additional Computationally Derived Assertions:** A number of additional computations contribute evidence including derived physical and chemical metrics such as signal peptide scores by SignalP (14), lipoprotein (LP) signals (16), transmembrane helices by TmHMM (15), molecular weight, isoelectric point, percent GC, outer membrane protein signals, PROSITE motifs (16), topological predictions and secondary structure predictions. Proteins are also searched against the InterPro member databases (e.g., SMART, SCOP, PRINTS) (17). While all searches are applied to all genes/proteins, with the exception of TmHMM and LP the other searches are not yet used in the assignment of automated annotation.

**4.4.6 AutoAnnotate:** AutoAnnotate is a programmatic approach to assigning functional annotation to gene models following JCVI's naming conventions guidelines (http://cmr.jcvi.org/CMR/NamingConventions.shtml) in an automated fashion. AutoAnnotate program uses a heuristic approach to evaluate results of homology searches, weighs the evidence using precedence-based rules that favor more trusted annotation sources to assign preliminary automated annotation to the gene models, including functional name, gene symbol, EC number, functional role category and Gene Ontology (GO) terms to each protein in the genome. Each protein is assigned a descriptive functional name coming from, in order of rank: the CHAR database, a trusted protein family, a best protein BLAST match from JCVI's PANDA non redundant protein database and computationally derived assertions.

## 5 Discussion

**N/A**

## 6 Related Documents & References

1. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997 Mar 1;25(5):955-64.

2. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007 Jan 19

3. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.

4. Nucleic Acids Research, 2006, Vol. 34, Database issue D32-D36

5. Proc Natl Acad Sci U S A 99, 2275 (Feb 19, 2002).

6. BMC Bioinformatics. 2007 Jan 20;8:18.

7. Nucleic Acids Research Advance Access first published online on May 30, 2007

8. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

9. The Enzyme Commission. *Enzyme Nomenclature* 1992 [Academic Press, San Diego, California, ISBN 0-12-227164-5 (hardback), 0-12-227165-3 (paperback)] with Supplement 1 (1993), Supplement 2 (1994), Supplement 3 (1995), Supplement 4 (1997) and Supplement 5 (in Eur. J. Biochem. 1994, 223, 1-5; Eur. J. Biochem. 1995, 232, 1-6; Eur. J. Biochem. 1996, 237, 1-5; Eur. J. Biochem. 1997, 250; 1-6, and Eur. J. Biochem. 1999, 264, 610-650; respectively)

10. The TIGRFAMs database of protein families. *Nucleic Acids Res*, 2003. 31(1): p. 371-3

11. Curr Protoc Bioinformatics. 2003 May;Chapter 2:Unit 2.5.

12. Profile hidden Markov models. Bioinformatics. 1998;14(9):755-63. Review.

13. Identification of common molecular subsequences. J Mol Biol. 1981 Mar 25;147(1):195-7.

14. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol., 340:783-795, 2004.

15. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001 Jan 19;305(3):567-80.

16. The PROSITE database, its status in 2002. Nucleic Acids Res. 2002 Jan 1;30(1):235-8.

17. Nucleic Acids Res. 2009 Jan;37(Database issue):D211-5. Epub 2008 Oct 21

18. http://cmr.jcvi.org/CMR/AnnotationSops.shtml
http://cmr.jcvi.org/CMR/NamingConventions.shtml

## 7 Revision History

**Author**: Ramana Madupu
**Version**: 1.01
**Effective Date**: 4/1/2009

| Version | Author/Reviewer | Date | Change Made |
|---|---|---|---|
| 1.01 | R. Madupu | 4/1/2009 | Establish SOP |