Common Gene Annotation Process
Broad Institute, WUGC, JCVI and Baylor

Author:
Version:
Effective Date:

1 Abstract

2 Introduction

This SOP describes the current annotation methodologies used by the four annotation centers in an effort to build a consensus for defining a set of minimum standards for annotation of HMP genomes, while preserving the distinctive strengths and innovation of each group. For further details about an individual centers pipeline, center specific SOP's can be found on the DACC at http://hmpdacc.org/tools_protocols/tools_protocols.php.

3 Requirements

4 Procedure

The procedure is broken down into two sections:

- **Structural annotation** – all 4 centers use their own best practices, described here, to perform structural annotation
- **Functional annotation** – the 4 centers have adopted the JCVI pipeline, described here, for consistency of functional annotation

4.1 Structural Annotation

4.1.1. Standard Computes

4.1.1.1. Blast

Blast homology to previously annotated proteins in the NR database provides useful information for the evaluation of ab initio predictions. In the absence of EST data for prokaryotes, blast data is the most useful resource for building high confidence gene models by the automated annotation pipelines. Blast parameters used by the different centers are shown in the chart below.

	Broad	**WUGSC**	**JCVI**	**BCM**
Blast Database	NR (bacteria)	NR (bacteria)	All Group NIAA-PANDA	NR (bacteria)
Min E value	10^{-10}	10^{-6} bit score=130	BlastP score cut off: 50 BlastP min E value: 0.1	10^{-5}
Min % identity	30%	30%	-	30%
Min query coverage	30%	30%	-	30%

Minimal standards for blast parameters were set to be inclusive of center specific pipelines. Work will continue to evaluate the use of bit scores in the pipelines to account for any variability in the database size, but moving forward the use of E values not greater than 10e-5 will be used and min % identity and coverage of 30% where applicable. The exception for these guidelines will be JCVI due to the unique nature of their blast databases and pipeline options that require personalized criteria. Specific details can be found in the JCVI-specific SOP at http://hmpdacc.org/tools_protocols/tools_protocols.php.

4.1.1.2. tRNAScan / Aragorn

tRNA are a key biological entity. The tRNA repertoire of an organism affects the codon bias seen in highly expressed protein coding genes.

- There is almost perfect agreement among all centers in the usage of tRNAscan to find tRNA features.
- Consensus on how tRNA information could be used to exclude spurious ORF predictions or prune over-extended predictions overlapping tRNA's can be found in section 4.1.3.5.

4.1.1.3. Rfam and RNAmmer

Rfam and RNAmmer predict common RNA features such as ribosomal RNA, small regulatory non-coding RNA, noncoding RNA. Besides representing these useful biological entities on the genome, the presence and organization of the features provide contextual information between RNAs and protein coding genes and further aid in the removal of spurious protein coding predictions. rRNA operons or clusters, if and when represented fully, are good indicators of the completeness of the genome assembly.

- All centers use similar options to exclude ORFs with overlaps to RNA features.

4.1.1.4. Hmmer

Pfam domains are very useful for assigning gene product names and for the functional annotation of genes. In addition, Pfam domains on a genomic axis could serve as useful landmarks for finding small protein-coding genes missed by the commonly used ab initio gene predictors. When present at complex loci with overlapping predictions, on both strands or in the same stand, they are very helpful in resolving overlaps resulting in the deletion of spurious predictions and selection of the best gene model.

	Broad	WUGSC	JCVI	BCM
Database	pFAM	pFAM		pFAM
LD Hmmer-PFAM	-	-	pFAM Library	-

LD Hmmer-TIGRFAM	-	-	TIGRFAM Library	-

4.1.2. Gene Finding

Find all potential protein coding genes on draft genome assemblies. Programs used: GLIMMER, GENEMARK, METAGENE etc. Gene finding programs use slightly different algorithmic and heuristic approaches for finding potential coding genes. This is obvious in terms of the observed differences in the gene count and their predicted structure by different programs.

	Broad	WUGSC	JCVI	BCM
Glimmer3	overlap=200 minLength=90 codonTable=11 linear	overlap=200 minLength=90 codonTable=11 linear	overlap=50* minLength=90 codonTable=11 linear	overlap=200 minLength=90 codonTable=11 linear
GeneMark	Genome Specific parameter file	Genome specific parameter file (build icm)	-	Genome Specific parameter file
MetaGene	Default	-	-	-
BER features	In house	-	repraze	-
pFAM ORFs	-	-	-	-
FgeneB	-	-	Default	-

*currently includes manual evaluation

- Different minimum ORF-length cut-offs used by different gene finding tools contribute to the differences in the gene numbers.
- Trained and untrained versions of the same ab initio prediction program on the genomes with high and low GC-rich genomes may produce slightly different predictions.

4.1.3. Gene Calling

Choosing the best genes from a population of ab initio and evidence-based gene models is the most important step in generating consistent gene models. This process includes:

- Generating both ab initio and evidence-based (blast and pFAM) predictions using one or more gene finding algorithms.
- Defining loci by clustering predictions with the same reading frame.
- Selecting the best of the predictions at each loci by evaluating them against the best evidence-blast and pFAM
- Picking a consensus gene model at each loci with only ab initio predictions
- Resolving overlaps between adjacent coding genes as well as non-coding features such as tRNAs and rRNAs.

Below is a table containing different cut-offs used by the four centers for defining the minimum ORF size with and without evidence.

1. *Minimum length of gene*

	Broad	WUGSC	JCVI	BCM
MinGeneLengthWithoutEvidence	120 bases	120 bases	90 bases	120 bases
MinGeneLengthWithEvidence	60 bases	60 bases	90 bases	60 bases

2. *Selecting best prediction at each loci*

	Broad	WUGSC	JCVI	BCM
Best Prediction selection	Best evidence-blast and pFAM	Choose GeneMark over Glimmer3 with same stop codon	Best evidence-blast and PFAM/TIGRFAM	Best evidence-blast and pFAM

3. *Resolving overlaps between 2 predictions*
Whenever two protein-coding ORFs with different reading frames overlap, each center implements a set of selection criteria. This criterion varies among the centers. In order to achieve greater uniformity, a consensus on how to resolve such overlaps was established and is shown in the charts below.

	Broad	WUGSC	JCVI	BCM
Maximum overlap allowed	200bp	200bp or 30% ORF Length evaluated	50bp	30% ORF length and not more than 200 bases

Resolving overlapping predictions
1. If both are predicted only, keep the longest orf
2. If both contain pFAM, keep both
3. If one has pFAM and blast and the other doesn't, keep the one with the pFAM hit
4. If one has pFAM and other low confidence blast, keep the one with the PFAM domain
5. If both have blast and pFAM, keep both (lesser overlapping in the silico prediction is chosen)
6. ORFs within ORFs- same or different strand and ORFs with the same reading frames are never allowed, even if both have evidence- choose the longest orf in this case.

4. *Conflict resolution in gene calls*
   - In case of overlapping predictions with different ORF lengths, blast evidence serves as a reference data point for picking the best gene models. In addition, blast evidence is also used for retaining overlaps between two adjacent genes.
   - Blast features are also used to create Blast extended features which are very useful for finding genes missed by commonly used ab initio predictors.

5. *Exclusion of open reading frames with overlaps to non-coding features*

	Broad	**WUGSC**	**JCVI**	**BCM**
Overlap to RNA features	Exclude unless a known gene	10-50% overlap allowed on same strand depending on rna type	Excluded unless a known gene	>30% length overlap on either strand is excluded

6. *Detection and tagging of ORFs with frame shifts*

   ORFs with one or more frame shifts are referred to as disrupted ORFs (dORFs). Disruption in an ORF may be caused by sequence errors or degeneration of the coding sequence leading to creation of pseudogenes. On finished genomes, these dORFs are referred to as pseudogenes. However on the draft genomes, one cannot easily characterize them as pseudogenes as they could be a result of common sequence problems and gaps in the assembled sequence.

   Disrupted ORFs can only be detected when they have blast evidence with indications of frame shift. At such loci, ab initio predictions will either split genes into two or more reading frames or predict a single ORF that is significantly shorter in length as compared to the evidence. Despite the fact that dORFs are common among bacteria, what makes the detection of these dORFs on draft genomes even more difficult is that not all split genes with blast evidence are dORFs: Some of them represent *real* splits. According to the rosette-stone hypothesis two or more functionally related genes can occur as either single ORFs or two or more smaller ORFs, each representing a functional ORF.

   The consensus among the genome centers is that we should tag the easily identifiable defective ORFs with the curation flag 'contains frame shift' to indicate the presence of the frame shift. Blast extended ORFs or genewise predictions are the common predictions capable of identifying dORFs.

   In general, the following issues in the blast extended ORFs indicate the presence of dORFs:
   - A single blast loci with two ab initio predictions in which each prediction corresponds to

a part of a single blast alignment.
- Two or more blast alignments in different reading frames.
- Only a fraction of the ORF is recognizable as compared to the blast query sequence.

4.2 Functional Annotation

The translated sequence of each gene model (identified by Glimmer or other methods) is searched against a variety of public and private databases. These search results are either stored in genome project specific databases, or maintained as protein specific search files retrievable for analysis.

4.2.1. Homology Searches

4.2.1.1. Non-identical amino acid database (NIAA)

Each protein is searched against an internal non-identical amino acid database (NIAA) comprised of all proteins available from GenBank (HYPERLINK "http://www.ncbi.nlm.nih.gov" http://www.ncbi.nlm.nih.gov), PDB (HYPERLINK "http://www.rcsb.org/pdb/Welcome.do" http://www.rcsb.org/pdb/ Welcome.do), UniProt (http://www.pir2.uniprot.org/), and the Comprehensive Microbial Resource database (http://www.tigr.org/CMR).

4.2.1.2. Blast-Extend-Repraze (BER)

The BLAST-Extend-Repraze (BER) search algorithm (HYPERLINK "http://ber.sourceforge.net" http://ber.sourceforge.net) initially runs a BLAST search (Altschul, et al., 1990) for each protein against NIAA and stores all significant matches in a mini-database. The nucleotide sequence of each gene is then extended 300nt upstream and downstream, and a modified Smith Waterman alignment (Smith et al., 1981) is performed against the mini-database. The extension of the sequence allows the resulting alignments to be evaluated for frameshift mutation or point mutations that introduce in-frame stop codons. If significant homology to a match protein exists and extends into a different frame from that predicted, or extends through a stop codon, the program continues the alignment past the boundaries of the predicted coding region.

4.2.2. Protein Families

4.2.2.1. Hidden Markov Models (HMMs)

Protein translations of each gene model are searched against Hidden Markov models (HMMs) using the HMMer package (Eddy, 1998) Two libraries of HMMs are used: the Pfam HMMs (Bateman, et al., 2000), and TIGRFAMs (Haft, et al., 2001).

4.2.3. Sequence signatures

Other amino acid sequence signatures, domains, or functional sites are predicted by searching all proteins against the PROSITE database (Falquet et al., 2002). The SignalP (Bendtsen et al., 2004) and TMHMM (Krogh et al., 2001) algorithms are used to predict putative signal sequences and membrane spanning domains respectively.

4.2.4. Gene Naming

AutoAnnotate is a programmatic approach to assigning descriptive functional annotation to gene models following JCVI naming convention guidelines in an automated fashion. It uses a heuristic approach to evaluate results of homology searches. By analyzing the BER and HMM search results, AutoAnnotate assigns a common name, gene symbol, Enzyme Commission (EC) number, and JCVI functional role categories and Gene Ontology terms (GO) as follows.

4.2.4.1.

AutoAnnotate first evaluates the isology and threshold score of each HMM match. If there is a hit to an equivalog-level HMM with a threshold score above the trusted cutoff score, the identifying information attached to that HMM (protein name, role category, gene symbol, GO terms and EC number if applicable) is assigned to the gene model.

4.2.4.2.

If there are no matches to an equivalog-level HMM with above the trusted cutoff score, the BER search results are evaluated. The program follows a specified ranking system to evaluate matches starting with the criteria to include a full-length match of at least 80% of the length of the subject and 80% identity. If more than one match is found, the program assigns highest rank to a Characterized protein from CHAR (**CHAR**acterized Protein Database). BER matches with varying percent identities over the length of the protein are ranked and prioritized in the following order: accessions from CHAR, Uniprot accessions and OMNI accessions - for annotation originating at JCVI. Interleaved into this ranking system are non-equivalog TIGRFAM and PFAM HMM hits to include subfamily, superfamily, domain and repeat HMM models. The specificity of final protein name is fine-tuned to reflect the evidence data type used to assign the protein name.

4.2.4.3.

Proteins with a pair-wise match to a hypothetical protein from another species, but no HMM hit, are named conserved hypothetical protein.

4.2.4.4.

Proteins with no HMM or BER matches remain named hypothetical protein.

4.2.4.5.

Hypothetical proteins with predicted lipoprotein signatures are names ‘putative, lipoprotein’. Hypothetical proteins with five or more predicted membrane spanning regions are names ‘putative membrane protein’.

5 Implementation

6 Discussion

7 Related Documents & References

8 Revision History

Version	Author/Reviewer	Date	Change Made
1.0			Establish SOP
1.0c		10/19/2011	Converted to standard template