

# Human Microbiome Project

## Annotation Summary from the 4 Collaborating Centers - Broad, WUGC, JCVI and Baylor

The purpose of this outline is to understand 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.

### SECTION – I STANDARD COMPUTES

#### 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.

	<b>Broad</b>	<b>WUGSC</b>	<b>JCVI</b>	<b>BCM</b>
<b>Blast Database</b>	NR (bacteria)	NR (bacteria)	All Group NIAA-PANDA	NR (bacteria)
<b>Min E value</b>	$10^{-10}$	$10^{-6}$ bit score=130	BlastP score cutt off : 50 BlastP min E value: 0.1	$10^{-5}$
<b>min % identity</b>	30%	30%	-	30%
<b>min query coverage</b>	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 which require personalized criteria. Specific details can be found in the JCVI SOP.

#### 2. tRNA scan / aragorn

tRNA's 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 III.5.

#### 3. Rfam and RNAmmer

Rfam and RNAmmer predict common RNA features such as ribosomal RNA, small regulatory non-coding RNA, non coding 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. 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
<b>Database</b>	pFAM	pFAM		pFAM
<b>LD Hmmer-PFAM</b>	-	-	pFAM library	-
<b>LD Hmmer-TIGRFAM</b>	-	-	TIGRFAM library	-

## SECTION – II 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
<b>Glimmer3</b>	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
<b>GeneMark</b>	Genome Specific parameter file	Genome specific parameter file (build icm)	-	Genome Specific parameter file
<b>MetaGene</b>	Default	-	-	-
<b>BER features</b>	In house	-	repraze	-
<b>pFAM ORFs</b>	-	-	-	-
<b>FgeneB</b>	-	-	Default	-

\*currently includes manual evaluation

- Different minimum ORF-length cut-offs used by different gene finding tools contribute to differences in the gene numbers.

- Trained and untrained versions of the same ab initio prediction program on genomes with high and low GC-rich genomes may produce slightly different predictions.

## SECTION – III 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

	<b>Broad</b>	<b>WUGSC</b>	<b>JCVI</b>	<b>BCM</b>
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

	<b>Broad</b>	<b>WUGSC</b>	<b>JCVI</b>	<b>BCM</b>
<b>Best Prediction selection</b>	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 it shown in the charts below.

	Broad	WUGSC	JCVI	BCM
<b>Maximum overlap allowed</b>	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 & blast and other doesn't, keep the one with 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 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
<b>Overlap to RNA features</b>	Excluded 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 can not

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.