

# ***Ixodes* MSC, UMD and VectorBase Project Plan**

The National Institute of Allergy and Infectious Diseases, National Institutes of Health has funded the *Ixodes scapularis* genome project through its Microbial Sequencing Centers (MSCs) at The J. Craig Venter Institute (JCVI) and the Broad Institute, Massachusetts Institute of Technology. The MSCs are responsible for genome sequencing, assembly, and annotation of gene structure and function, with the goal of rapid release of each of these data sets to the scientific community. Once released, the complete sequence and annotation of the *Ixodes* genome will permanently reside at a third NIAID-sponsored entity, VectorBase, which is a Bioinformatics Resource Center (BRC). Delivery of this data may also require coordination with NCBI-GenBank.

Given the mutual interests of these organizations, the most effective approach to the initial release of *Ixodes* genomic data will be to work in close collaboration to produce an initial set of annotations, refine and improve the pipelines resident at each of the centers, and to generate data to be used for the analysis and publication. With the current contig coverage of 3.82X and a GC content of 45% in the assembly, the combined annotation efforts of the MSCs, UMD and VectorBase are critical to produce unified, best possible annotation for release to the scientific community. This document is a project plan providing the specific details for activities planned among the three NIAID-funded centers in coordination with UMD by way of a sub-contract to Dr. Nene, who continues to lead the efforts on the project along with his JCVI counterpart.

With several different parties involved, effective communication at a detailed level is paramount throughout the course of this project. JCVI, UMD, Broad and VectorBase shall agree on primary points of contact and mechanisms of communication, such as conference calls or email lists.

The primary annotation contacts at each of the different centers shall be:

1. Elisabet Caler (ecaler@jcv.org) at JCVI,
2. Valentina Difrancesco (vdifrancesco@niaid.nih.gov) at NIAID
3. Martin Hammond (mhammond@ebi.ac.uk) at VectorBase
4. Catherine Hill (hillca@purdue.edu) at Purdue University
5. Chinappa Kodira (ckodira@broad.mit.edu) at the Broad Institute, and
6. Vish Nene (vnene@som.umd.edu) at University of Maryland.

The PIs from JCVI, UMD, Broad and VectorBase shall always be involved in meetings and conference calls in which important decisions are to be made. Conference calls with NIAID to provide updates on the status of the project shall be made on a regular basis and are currently scheduled for every second Thursday of the month till projects end.

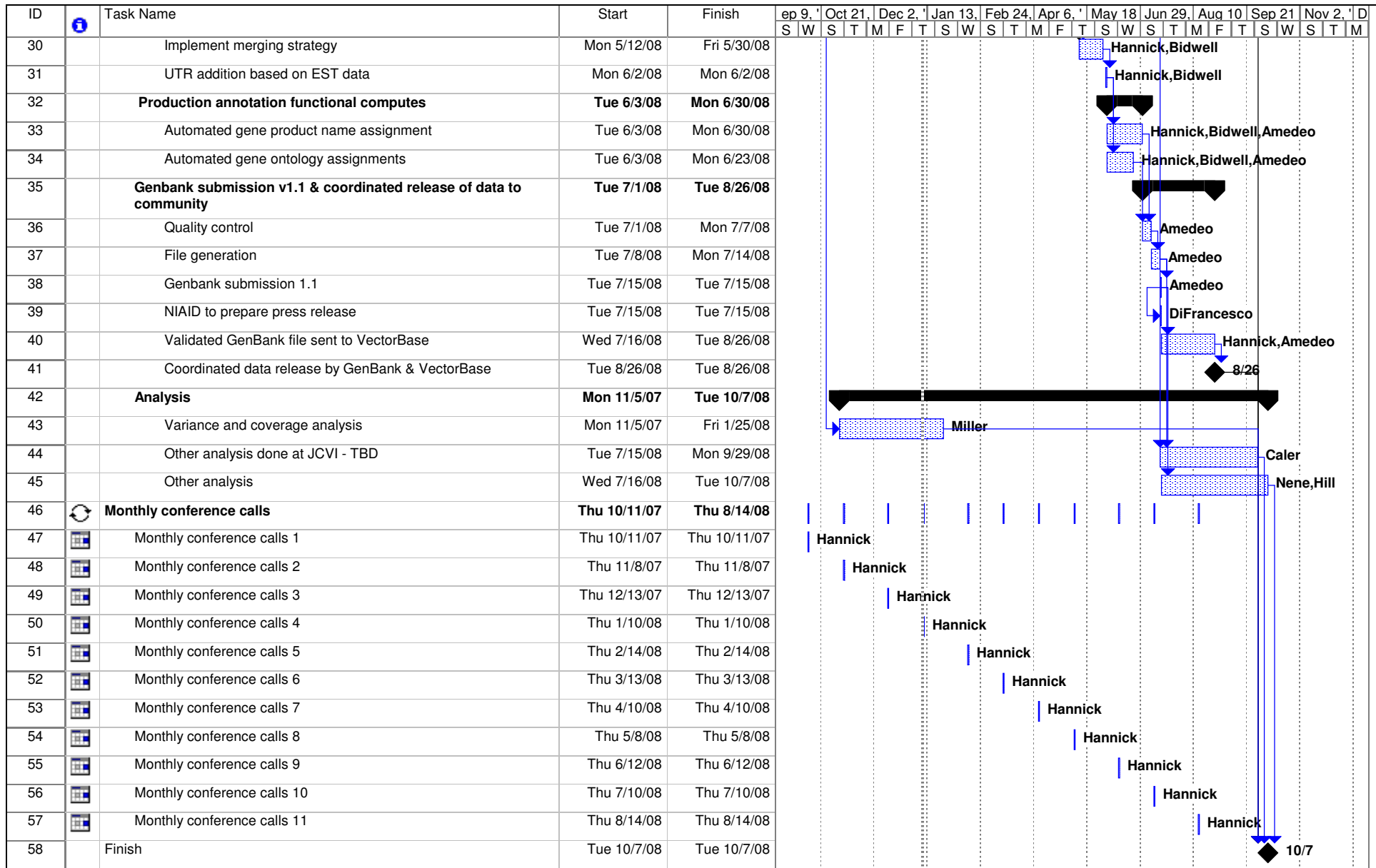
The milestones shall be reviewed, following input from Broad and VectorBase, with final review by NIAID. It is understood that if deadlines cannot be met or issues arise, we can review and re-visit the timelines and appropriately communicate to the larger project team. The designated representatives from each centers mentioned above as points of contacts shall serve as the project managers for the plan we put in place and be involved in communicating between centers.

Other points of agreement may also be required to achieve the milestones laid out in the timeline for GenBank submission of Release 1.0. The exact data types to be exchanged and the file formats associated with the data types need to be established. The metrics used for evaluation and comparison of the data sets produced by the Broad, VectorBase and JCVI shall also be articulated and shared.

Below is the project plan as developed jointly with all the participating groups. Descriptions of relevant tasks with specific details follow in text format thereafter.

ID	Task Name	Start	Finish	ep 9.	Oct 21.	Dec 2.	Jan 13.	Feb 24.	Apr 6.	May 18.	Jun 29.	Aug 10.	Sep 21.	Nov 2.	D			
				S	W	S	T	M	F	T	S	W	S	T	M	F	T	S
1	<b>Ixodes Project Start</b>	<b>Mon 11/5/07</b>	<b>Tue 10/7/08</b>															
2	<b>Ixodes assembly submission &amp; completion</b>	<b>Mon 11/5/07</b>	<b>Fri 12/14/07</b>															
3	Establish assembly release logistics with GenBank	Mon 11/5/07	Fri 12/14/07															
4	WGS submission to GenBank	Mon 11/5/07	Fri 12/14/07															
5	Assembly submission complete	Fri 12/14/07	Fri 12/14/07															
6	<b>Annotation preparation</b>	<b>Mon 11/5/07</b>	<b>Fri 12/14/07</b>															
7	Create central repository (CR)	Mon 11/5/07	Mon 11/5/07															
8	Exchange of useful datasets (repeat libraries etc.)	Tue 11/6/07	Mon 11/26/07															
9	Define annotation data types and metrics for evaluating gene sets	Mon 11/5/07	Fri 12/14/07															
10	<b>Production annotation gene structure</b>	<b>Mon 11/5/07</b>	<b>Tue 4/8/08</b>															
11	Run autopipeline on all scaffolds > 10Kb	Tue 11/27/07	Mon 2/18/08															
12	Generate JCVI 0.5 preliminary annotation	Mon 2/18/08	Mon 2/18/08															
13	<b>Iterative improvement of gene set</b>	<b>Tue 2/19/08</b>	<b>Mon 3/31/08</b>															
14	Verify against relevant data sets	Tue 2/19/08	Mon 3/3/08															
15	Locate genes in introns of others	Tue 3/4/08	Mon 3/17/08															
16	EST data incorporation	Tue 3/18/08	Mon 3/31/08															
17	Quality control	Tue 4/1/08	Mon 4/7/08															
18	Submit JCVI 0.5 in central repository	Mon 4/7/08	Mon 4/7/08															
19	<b>Vectorbase: gene structure annotation</b>	<b>Mon 11/5/07</b>	<b>Tue 4/8/08</b>															
20	Vectorbase annotation	Mon 11/5/07	Fri 3/21/08															
21	Submit VB v0.5 into central repository	Fri 3/21/08	Fri 3/21/08															
22	Make v 0.5 available via web	Tue 4/8/08	Tue 4/8/08															
23	<b>Broad: annotation</b>	<b>Mon 11/5/07</b>	<b>Fri 12/28/07</b>															
24	Develop training sets	Mon 11/5/07	Fri 12/28/07															
25	Submit Broad data in central repository	Fri 12/28/07	Fri 12/28/07															
26	<b>Evaluation of data</b>	<b>Wed 4/9/08</b>	<b>Fri 5/9/08</b>															
27	Evaluate and compare gene sets	Wed 4/9/08	Tue 5/6/08															
28	Define merge strategy based on evaluation (done via a meeting)	Wed 5/7/08	Fri 5/9/08															
29	<b>Data generation gene structure v1.0</b>	<b>Mon 5/12/08</b>	<b>Mon 6/2/08</b>															

Project: Ixodes Date: Wed 1/9/08	Task		Milestone		External Tasks	
	Split		Summary		External Milestone	
	Progress		Project Summary		Deadline	



Project: Ixodes Date: Wed 1/9/08	Task		Milestone		External Tasks	
	Split		Summary		External Milestone	
	Progress		Project Summary		Deadline	

### **Task 3: Develop assembly release strategy**

JCVI shall make the WGS submission and work with GenBank to ensure VectorBase is represented in the submission so they can provide subsequent updates to the gene set.

Text describing the Ixodes assembly will have to be added to the GenBank record. Such text will be reviewed by NIAID prior to the release of the GenBank record.

The centers will together define the assembly data types and the nomenclature to be used that will be submitted to GenBank, and the divisions to which they will be submitted. These submissions will be documented to describe how the assembly was produced, general statistics about the assembly, and information about future releases. Broad and JCVI will coordinate public web access to the assembly data through VectorBase. This data access will include file downloads and BLAST search capability. Decisions for additional assembly related analysis shall be completed and communicated to the larger team.

Preparation for production annotation will occur at this stage. Data sets that are assumed to be useful to the project will also be stored in a central repository for ftp exchange.

### **Task 7: Create central repository (CR)**

JCVI's IT department shall create an FTP directory with password protection if not already done. The purpose of this site is to serve as a central repository for temporary data to allow for exchanges between the Broad, JCVI, UMD and VectorBase.

### **Task 8: Exchange useful datasets**

It is anticipated during the course of the project that there shall be a need to share pre-annotation data sets that will be useful for the annotation efforts. This data shall be considered work products and not released to the scientific community. The information shall be stored in a central password protected FTP site. Readme files shall accompany the data files.

JCVI has produced and shall make available the following data sets:

- Repeat library of Ixodes-specific repeats
- Multi-species transposon ORF database
- Repeat masked genomic sequence
- Assembled EST sequences based on genome alignment and clustering
- Gene prediction program output

### **Task 9: Define annotation data types and metrics for evaluating gene sets**

**annotation data types and file formats:** The data types for the merging of data shall be CDS of protein coding genes. The file formats for exchange of data, particularly for the data that will be production Annotation gene structure will be GFF and Fasta file formats.

The data types proposed for GenBank Submission v1.0 are: protein coding genes + UTRs, gene product names, and automated Gene Ontology assignments. The following

secondary Data types, where these data types will be submitted if available but shall not become sources of delays to production of Release 1.0: alternative splice forms, transposons, repeat annotation, and non-coding RNAs.

**Metrics for evaluating gene sets:**

*The parties that will participate in the evaluation and decision making strategy for merge must be established.* The specific metrics for the merge strategy will also be established.

For each gene set generated by the centers JCVI shall:

- Evaluate each gene set against the available EST data
- Evaluate each gene set against relevant comparative genomics data (Aedes, Drosophila)
- Compare gene sets for identical, overlapping and unique genes

**Task 11: Run autopipeline (assembly level computes) on all scaffolds > 10Kb**

Preliminary results have already shown that optimizing the pipeline for Ixodes is once again going to be an R&D effort similar to that of Aedes and Culex, requiring multiple iterations of parameterization/training and evaluation to arrive at the best possible gene set.

Thus, the major challenges facing us during this stage are definition of correct gene boundaries, dealing with partial genes caused by the fragmented assembly and comprehensive repeat identification and screening.

The data generated by the centers in this stage shall be deposited in the central repository for use in the evaluation of data, and lead to an active exchange of methods and data between JCVI, the Broad Institute, VectorBase and UMD as further refinement of the dataset progresses.

JCVI's annotation pipeline is a flexible, robust framework that has been used to annotate diverse eukaryotic genomes, from the relatively small genomes of protists and fungi, to the larger and more complex genomes of plants, nematodes and insects. It establishes a workflow around a series of software packages for gene prediction, repeat identification, and nucleotide and protein alignments.

We will train and evaluate the following gene finders on the training sets generated by JCVI and Broad:

Ab Initio:

- GeneZilla (Majoros et al. Bioinformatics. 2004:2878-9)
- Genie (Reese et al. Genome Res. 2000 :529-38)
- AUGUSTUS (Stanke et al. Nucleic Acids Res. 2004:32)

Homology-based:

- Twinscan (Korf et al. Bioinformatics. Suppl 1:S140-8) using an informant genome
- Genewise

The basic pipeline for gene structure annotation consists of the best-performing of the gene finders listed above as well as a set of similarity-based computes. The datasets include a non-redundant amino acid database filtered from public sources, a dataset of insect proteins parsed from the above, and sequences representative of PFAM profiles.

Gene models shall be generated from the above along with those generated from VectorBase using Evidence Modeler, which synthesizes protein and EST alignments with the data from all the gene predictors. PASA (Program to Assemble Spliced Alignments; Haas et al. Nucleic Acids Res. 2003:5654-66) will be used to ensure that the gene models are consistent with available EST and cDNA information.

#### **Task 14: Verify against relevant data sets and iteratively improve the gene set**

JCVI 0.5 shall represent the raw output of the JCVI pipeline, a set of consensus gene predictions produced by a program called Evidence Modeler, which synthesize protein and EST alignments with the data from the individual gene prediction programs.

Due to the complex nature of this genome project, we anticipate the need to process iterative improvements to the initial data set. We shall identify missed and misannotated genes by checking 0.5 predictions against the protein alignments from *Aedes aegypti*, *Culex pipiens*, Drosophila and other relevant genomes available at this time, screening introns larger than 10kb for the presence of possible protein coding genes, and by comparing the predictions to the aligned and assembled ESTs.

These data shall then be reviewed, and gene structures added or modified computationally based on an appropriate set of criteria to be determined.

#### **Task 15: Locate genes in introns of others**

Most gene prediction tools do not correctly predict nested and overlapping genes. During the pre-annotation analysis, if like in Aedes, examples of protein coding genes in the introns of other protein coding genes are identified then, we shall extract all intron sequences greater than 10kb, mask for repeats, and process them separately to capture gene predictions that may have been missed.

#### **Task 16: EST data incorporation**

Once we are confident that the gene set is substantively complete, the PASA program shall be used to compare the gene set with the aligned and assembled EST sequences. Conflicts reported by PASA between EST alignments and the gene structures shall be processed in an automated fashion to resolve the identified differences. PASA reported intergenic EST alignments shall be analyzed for homology to transposons, presence of an open reading frame and length to determine if additional genes should be added before finalizing the gene set.

## **Task 17: Quality control**

Quality control consists of a set of data integrity and quality checks meant to ensure that the data is as error free and accurate as possible.

Some of these QC steps will result in the elimination or editing of existing gene models, others will result in a standard comment appended to the gene.

These include but are not limited to:

- Presence of start codon
- Presence of stop codon
- Presence of consensus splice sites
- Intron length above minimum
- Exon length above minimum
- Protein length above minimum
- Exon coordinates map within gene coordinates
- Exons do not include Ns

Based on preliminary results, we also anticipate that there may be over prediction of genes. We may, as part of this QC process, screen again for transposon contamination and eliminate or note smaller proteins with no protein or EST evidence.

Gene level QC may also necessitate reviewing the assembly data, including underlying reads, to assess sequence quality and its effect on the annotation pipeline output.

## **Task 18: Submit JCVI 0.5 in central repository**

The subsequent improvements to JCVI's gene set shall be deposited in Central Repository FTP site.

## **Task 20: VectorBase autopipeline**

The VectorBase analysis team at Ensembl has an automated gene build process which has been deployed very successfully on a broad range of vertebrates and invertebrates.

Ensembl annotates from evidence, either cDNAs or ESTs from the organism itself or protein sequences from related organisms. The genome sequences are repeatmasked and gene sets from different approaches are prepared.

There are five main sources genes:

Set 1. Models built with Genewise using species-specific proteins (from UniProt + community-contributed models where available), given EST/ mRNA-based extensions if possible, and merged to produce a non- redundant set.

Set 2. Models built with Genewise using high-quality arthropod proteins from UniProt, and merged to produce a non-redundant set.

Set 3. Models built solely from Ixodes ESTs using an in-house merging system.

Set 4. Models built with Genewise using other Metazoan proteins from UniProt, and merged to produce a non-redundant set.

Set 5. SNAP *ab initio* predictions that have an identifiable Pfam domain - rigorously screened to try to avoid TE-based predictions.

These five sources of genes are then merged into one final set, such that the highest confidence set is built on by progressively adding non-overlapping models from lower confidence sets. Prioritizing Sets 2/3/4 varies between species - for *Ixodes* the emphasis on ESTs would be increased to evaluate if some more sophisticated merging of Sets 2 and 3, rather than treating them as separate sets.

### **Task 21: Submit VectorBase 0.5 into central repository**

The results of the VectorBase annotation pipeline shall be deposited in Central Repository FTP site.

### **Task 22: make v 0.5 available via web**

VectorBase shall produce an EnsEMBL pre-site for *Ixodes*. An EnsEMBL pre-site gives access to the assembly for a genome using the same 'look and feel' of the EnsEMBL browser. A first pass set of gene structures is available for browsing and searching. All EnsEMBL genes are evidenced by protein or cDNA data and hence we can provide a nomenclature which both is informative as to how the gene prediction was produced (in terms of similarity) and yet is obviously a work in progress. There is no attempt to track or version such predictions.

We envisage users to access the site via the BLAST server which leads them into the browser to investigate their region of interest in greater detail. This aspect of a pre-site relies on the ability to run a BLAST/SSAHA search of the assembled genome, a repeatmasked version of the assembly and predicted cDNA/peptide data sets.

The pre-site is planned to contain the following data:

Sequence data:

Genome sequence and associated assembly information (from the supplied AGP file)

Similarity data:

cDNA/EST transcripts aligned using exonerate.

UniProt proteins aligned using Blast.

Gene predictions:

0.5 releases from JCVI and VectorBase.

*ab initio* predictors (Augustus, GeneZilla, Genie, and SNAP).

Homology predictors (Genewise, Twinscan)

Gene prediction sets will be uploaded to the central repository FTP site in GFF format and VectorBase will be responsible for converting these into EnsEMBL format for inclusion on the pre-site.

### **Task 24: Broad annotation**

Provide training sets for inclusion

### **Task 25: Submit Broad annotation in central repository**

The results of the Broad Center's inputs shall be deposited in Central Repository FTP site.

### **Task 27: Evaluating the gene set**

The goal of the structural annotation effort is the production of a single, high-quality set of gene predictions for release to the scientific community that shall be deposited into GenBank.

Having access to the annotations via a genome browser as a result of task 22, VectorBase can use the existing community annotation system to store annotations. The system is based on a CHADO schema relational database which stores the predictions with author information and relevant supporting evidence. Predictions shall be uploaded to the database using a web-based submission procedure involving Excel spreadsheets. Submission of a new annotation triggers a process by which the data is appraised by a human annotator (in the case of *Ixodes* this will be the community representative Jason Meyer, based in Cate Hill's laboratory in Purdue University). Once approved by Jason the prediction shall be displayed on the genome browser via DAS (Distributed Annotation System). The process of human quality control at this stage shall allow us to have high confidence in these annotations when it comes to integration with the JCVI/VectorBase geneset's.

Annotations collated from the community (and any other interested parties) shall be stored in a relational database. VectorBase shall extract this data and transform it into a number of file formats (mainly GFF3). The proposal is to integrate these annotations with the 1.0 merged set as part of the refinement and quality control process prior to the generation of the final 1.1 gene set for release (submission to GenBank). We shall strive to obtain corrected gene structures and assignment of established gene symbols during this process.

As the process of collating and processing the community driven manual annotations maybe time consuming, it is suggested that a deadline be set to coincide with the generation of the 1.0 gene set. The listserv in use amongst the *Ixodes* community shall facilitate and expedite communications as well. At this point the approved data in the community annotation database will be parsed into a GFF3 file and sent to JCVI for integration. Any further community involvement will be processed by VectorBase and integrated into a future gene set after submission to GenBank and publication.

### **Task 28: Defining selection strategy based on evaluation**

All centers shall review the results of data evaluation with input from the *Ixodes* community and come to a decision on the composition of the final gene set. Based on our

experience with *Aedes* and *Culex* projects, this process was rather involved and was successfully completed with face-to-face meetings. We therefore recommend a similar meeting be organized at the JCVI to finalize the merge strategy for *Ixodes* as well.

### **Task 30: Generate the final gene set at JCVI/Broad**

Once the strategy for developing a single, robust gene set for GenBank release is agreed upon, JCVI shall implement the strategy, leveraging existing tools for the automated update of gene models, assess and resolve any remaining differences between data sets, perform quality assurance on gene predictions, run functional computes and release the data to the public.

### **Task 31: UTR addition based on EST data**

Once a set of gene structures is established, the PASA pipeline shall be used to automatically attach UTR features where suggested by the EST evidence.

### **Task 33: Automated gene product name assignment**

In order to make the gene structure information usable by the scientific community, gene product names and Gene Ontology assignments shall be attached to each gene model computationally.

Once gene models are created, the JCVI annotation group shall run a series of protein function computes to assist in the identification of pathways and families of interest for publication. These computes include:

- Search of Panther, Experimentally-verified Panda, Drosophila and Panda (general) using BLASTP to identify top hit
- Pfam search using HMMER2
- Search of PROSITE, PRINTS, and ProDom, followed by Interpro classification including the results from the Pfam searches using InterProScan
- Transmembrane domain identification using TMHMM
- Signal peptide prediction using SignalP
- Protein families calculation

Gene product names shall be assigned based on significant domain and BLASTP matches. In order to organize the annotation data for further analysis, proteins will be organized into family groupings based on linkage clustering or conserved domain composition.

### **Task 34: Automated Gene Ontology assignments**

Gene products will also be assigned to Gene Ontology (GO) terms by transferring the GO associations of the best *Drosophila melanogaster* protein match and/or by using the mappings between PFAM/InterPro and Gene Ontology terms.

### **Task 36: Quality control**

Quality control consists of a set of data integrity and quality checks meant to ensure that the data is as error free and accurate as possible. In addition to final checks of gene structure data gene name and Gene Ontology information for data and formatting problems shall also be verified.

### **Task 37: File generation**

The data for Release 1.0 is prepared for submission into GenBank. The data conveyed shall be the data types and shall include: protein coding genes + UTRs, gene product names, and automated Gene Ontology assignments. We shall also attempt to submit the following secondary data types, where these data types will be submitted if available but will not become sources of delays to production of Release 1.0: alternative splice forms, transposons, repeat annotation, and non-coding RNAs. As mentioned earlier nomenclature used shall conform to VectorBase practice. All data files shall be shared internally prior to GenBank submission.

### **Task 38: GenBank submission 1.0**

At this stage the centers will have coordinated the production of a final, unified data set of gene predictions that reflect the contributed data. The sequences, gene structures, and functional annotations of Release 1.0 will be deposited into GenBank. The data for Release 1.0 is submitted into GenBank. The submission will address coordination of records in GenBank's WGS section in context of its other sequence submission areas. On GenBank processing to release from the NCBI web site:

*WGS projects without annotation require at least two weeks to be processed. Projects with annotation require at least one month for processing. Please submit your project with enough lead time.*

### **Task 39: NIAID to prepare the press release**

Press releases shall be prepared by the NIAID with input provided by the sequencing centers and UMD during the time that the GenBank record is processed by NCBI staff. Statistics on the Ixodes genome shall also be made available.

### **Task 40: Validated GenBank files sent to VectorBase for processing**

Once the submitted GenBank files representing the 1.0 annotation release pass the standard validation criteria and are reviewed and accepted by GenBank, these files shall be made available to VectorBase for processing so that the GenBank and VectorBase releases are synchronized.

### **Task 41: Data release/press release**

Press statements on the availability of the Ixodes genome are released, the data appears at NCBI. Data will also appear on ftp and web sites of the consortium.