

Aedes MSC and VectorBase Project Plan

The National Institute of Allergy and Infectious Diseases, National Institutes of Health has funded the *Aedes aegypti* genome project through its Microbial Sequencing Centers (MSCs) at The Institute for Genomic Research (TIGR) and the Broad Institute. 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 *Aedes* genome will permanently reside at a third NIAID-sponsored entity, VectorBase, which is a Bioinformatics Resource Center (BRC) at the University Of Notre Dame. Delivery of these data will also require coordination with NCBI-GenBank. Given the mutual interests of these organizations, the most effective approach to the initial release of *Aedes* genomic data will be to work in close collaboration to produce an initial set of annotation, refine and improve the pipelines resident at each of the centers, and to generate data to use in the analysis and publication of *Aedes*.

We are confident that the combined annotation efforts of the MSCs and VectorBase produce unified, high-quality *Aedes* annotation for release to the scientific community. This document is a detail project plan intended to describe the complementary activities planned among the three NIAID-funded centers.

See the accompanying document [Timeline.pdf](#) for the task timeline overview.

A summary of the activities for the MSC *Aedes* project listed by institution

Institution: Broad			
Task	Duration	Start	Finish
Stage 0: Establish Communication Mechanisms Among TIGR, Broad and VectorBase	5 weeks	9/12/05	10/14/05
Develop Assembly Release Strategy	3 weeks	9/12/05	9/30/05
Conference Call	1 day	9/27/05	9/27/05
Review and Acceptance of Project Plan	2.4 weeks	9/22/05	10/07/05
Stage 1: Annotation Preparation	4 weeks	9/19/05	10/14/05
Release August Assembly	5 days	10/03/05	10/07/05
Release useful datasets	5 days	10/10/05	10/14/05
Run and Evaluate Computes	3 weeks	9/26/05	10/14/05
Stage 2: Production Annotation Gene Structure	8 weeks	10/17/05	12/09/05

Broad: 10 Mb annotation	8 weeks	10/17/05	12/09/05
Define and Annotate 10 Mb region	8 weeks	10/17/05	12/09/05
Submit 10 Mb annotation in Central Repository	1 day	12/09/05	12/09/05
Stage 3: Evaluation of Data	4 weeks	12/12/05	1/06/06
define merge strategy based on evaluation	5 days	1/02/06	1/06/06
Possibility of Holiday Related Delay	1 day	12/30/05	12/30/05
Stage 7: Genbank Processing to Release	4 weeks	3/10/06	4/07/06
Press Release	1 day	3/10/06	3/10/06
Broad	4 weeks	3/13/06	4/07/06
Stage 8: Manuscript Preparation	29 weeks	1/09/06	7/28/06
Define Publication Strategy	5 weeks	1/09/06	2/10/06
Data Analysis for Publication	20 weeks	2/13/06	6/30/06
Support for Collaborator Analysis	20 weeks	2/13/06	6/30/06

Institution: TIGR

Task	Duration	Start	Finish
Stage 0: Establish Communication Mechanisms Among TIGR, Broad and VectorBase	5 weeks	9/12/05	10/14/05
Develop Assembly Release Strategy	3 weeks	9/12/05	9/30/05
Develop Draft Project Plan for Annotation and Analysis	1 week	9/12/05	9/21/05
Define Annotation Data types and Metrics for Evaluating Gene Sets	3 weeks	9/12/05	9/30/05
Conference Call	1 day	9/27/05	9/27/05
Review and Acceptance of Project Plan	2.4 weeks	9/22/05	10/07/05
Stage 1: Annotation Preparation	4 weeks	9/19/05	10/14/05
Create Central Repository (CR)	2 weeks	9/19/05	9/30/05
Release useful datasets	5 days	10/10/05	10/14/05
Stage 2: Production Annotation Gene Structure	8 weeks	10/17/05	12/09/05
TIGR gene structure annotation	8 weeks	10/17/05	12/09/05
Run autopipeline	4 weeks	10/17/05	11/11/05
Submit TIGR 0.5 in Central Repository	1 day	11/11/05	11/11/05
Iterative improvement of gene set	3 weeks	11/14/05	12/02/05
Verify against Anopheles and Drosophila	5 days	11/14/05	11/18/05
Locate genes in introns of others	5 days	11/21/05	11/25/05
EST Data Incorporation	5 days	11/28/05	12/02/05
Quality Control	5 days	12/05/05	12/09/05

Submit 0.5.1 in Central Repository	1 day	12/09/05	12/09/05
Stage 3: Evaluation of Data	4 weeks	12/12/05	1/06/06
implement work in Proposal for evaluating and comparing gene sets	3 weeks	12/12/05	12/30/05
define merge strategy based on evaluation	5 days	1/02/06	1/06/06
Possibility of Holiday Related Delay	1 day	12/30/05	12/30/05
Stage 4: Data Generation Gene Structure v1.0	4 weeks	1/09/06	2/03/06
Implement Merging Strategy	3.6 weeks	1/09/06	2/01/06
UTR Addition based on EST data	2 days	2/02/06	2/03/06
Possibility of Annotation Tracking to New Assembly	1 day	2/03/06	2/03/06
Stage 5: Production Annotation Functional Computes	3 weeks	2/06/06	2/24/06
Automated Gene Product Name Assignment	3 weeks	2/06/06	2/24/06
Automated Gene Ontology Assignments	3 weeks	2/06/06	2/24/06
Stage 6: Genbank Submission v1.0	2 weeks	2/27/06	3/10/06
Quality Control	5 days	2/27/06	3/03/06
File Generation	5 days	3/06/06	3/10/06
Genbank Submission 1.0	1 day	3/10/06	3/10/06
Stage 7: Genbank Processing to Release	4 weeks	3/10/06	4/07/06
Press Release	1 day	3/10/06	3/10/06
Broad	4 weeks	3/13/06	4/07/06
Stage 8: Manuscript Preparation	29 weeks	1/09/06	7/28/06
Define Publication Strategy	5 weeks	1/09/06	2/10/06
Data Analysis for Publication	20 weeks	2/13/06	6/30/06
Support for Collaborator Analysis	20 weeks	2/13/06	6/30/06

Institution: VectorBase

Task	Duration	Start	Finish
Stage 0: Establish Communication Mechanisms Among TIGR, Broad and VectorBase	5 weeks	9/12/05	10/14/05
Conference Call	1 day	9/27/05	9/27/05
Review and Acceptance of Project Plan	2.4 weeks	9/22/05	10/07/05
Stage 1: Annotation Preparation	4 weeks	9/19/05	10/14/05
Release useful datasets	5 days	10/10/05	10/14/05
Stage 2: Production Annotation Gene Structure	8 weeks	10/17/05	12/09/05
VectorBase: gene structure annotation	8 weeks	10/17/05	12/09/05

autopipeline	4 weeks	10/17/05	11/11/05
make VB v 0.5 available via web	4 weeks	11/14/05	12/09/05
Submit 0.5 or 0.5+ into Central Repository	1 day	12/09/05	12/09/05
Stage 3: Evaluation of Data	4 weeks	12/12/05	1/06/06
define merge strategy based on evaluation	5 days	1/02/06	1/06/06
Possibility of Holiday Related Delay	1 day	12/30/05	12/30/05
Stage 7: Genbank Processing to Release	4 weeks	3/10/06	4/07/06
Press Release	1 day	3/10/06	3/10/06
Broad	4 weeks	3/13/06	4/07/06
Stage 8: Manuscript Preparation	29 weeks	1/09/06	7/28/06
Define Publication Strategy	5 weeks	1/09/06	2/10/06
Data Analysis for Publication	20 weeks	2/13/06	6/30/06
Support for Collaborator Analysis	20 weeks	2/13/06	6/30/06

A summary of the Stages of completion for the MSC *Aedes* project

Stage 0: Establish Communication Mechanisms among TIGR, Broad and VectorBase (5 weeks)

Stage 1: Annotation Preparation (4 weeks)

Stage 2: Production Annotation Gene Structure (8 weeks)

Stage 3: Evaluation of Data (4 weeks)

Stage 4: Data Generation Gene Structure v1.0 (4 weeks)

Stage 5: Production Annotation Functional Computes (3 weeks)

Stage 6: Genbank Submission v1.0 (2 weeks)

Stage 7: Genbank Processing to Release (4 weeks)

Stage 8: Manuscript Preparation (29 weeks)

Stage 0: Establish Communication Mechanisms among TIGR, Broad and VectorBase (5 weeks)

Effective communication at this level of detail is paramount throughout the course of this project. Early in Stage 1, TIGR, Broad and VectorBase will agree on primary points of contact and mechanisms of communication, such as conference calls or email lists.

The primary annotation contacts are Jennifer Wortman (jwortman@tigr.org) and Vish Nene (nene@tigr.org) at TIGR, James Galagan (jgalag@broad.mit.edu) at the Broad Institute, and Dan Lawson (lawson@ebi.ac.uk) at VectorBase. The PIs from TIGR, Broad and VectorBase should also be involved in meetings and conference calls in which important decisions are to be made.

Conference calls with NIAID to update on the status of the annotation project should be made on a regular basis (monthly).

NIAID requested the development of a project management plan (this document) by October 1, 2005. The purpose of the project plan is to define the next steps of the project including the final assembly, annotation and analysis, and publication with milestones and timelines, including timelines for data releases. This will allow for better management of the project and coordination of effort between centers.

A final document will be created, following input from Broad and VectorBase, with final review by NIAID. It will be understood that if deadlines cannot be met or issues evolve, we can review and re-visit the milestones and timelines when needed. There will be designated representatives from each center who will serve as the project managers for the plan we put in place and be involved in communicating between centers.

Other points of agreement will 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 during the Production Annotation Stage and the file formats associated with those data types will be established. The metrics that will be used for evaluation and comparison of the gene sets produced by the Broad, VectorBase and TIGR will also be initiated during this phase of the project.

Task: Develop Assembly Release Strategy

Institutions: Broad, TIGR

Duration: 3 weeks

Activity: Broad and TIGR will work with Genbank to clarify co-ownership of WGS submissions to ensure that Broad can submit under the current TIGR WGS submission (Accession AAGE), and to ensure that TIGR can add the primary annotation to the Broad

assembly. VectorBase should also be represented in the submission so they can update the gene set post publication.

Text describing the Aedes 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 define the assembly data types 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 TIGR will coordinate public web access to the assembly data through their respective web sites. This data access will include file downloads and BLAST search capability.

Task: Develop Draft Project Plan for Annotation and Analysis

Institution: TIGR

Duration: 1 week

Activity: The initial draft of the project plan covering the final assembly, annotation and analysis, and preparation for publication is created in Microsoft Project and converted to a Microsoft Word document for circulation. The timeframe represented in the draft plan is the present (September 2005) through submission of a publication (estimated to be July 2006), with a projected annotation release date of March 2006.

Task: Define Annotation Data types and Metrics for Evaluating Gene Sets

Institution: TIGR

Duration: 3 weeks

Activity: Annotation Data types and File Formats

The data types for the merging of data in Stage 3 will be CDS of protein coding genes. The file formats for exchange of data, particularly for the data that will be used for Stage 3: Production Annotation Gene Structure will be GFF and Fasta file formats.

The data types proposed for Stage 7: 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 will 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 in Stage 3 must be established.

The specific metrics for the merge strategy will also be established.

For each gene set generated by the centers TIGR will:

- Evaluate each gene set against the available EST data
- Evaluate each gene set against comparative genomics data (Anopheles, Drosophila)
- Evaluate each gene set against the manually curated 10 Mb from Broad
- Compare gene sets for identical, overlapping and unique genes

Task: Conference Call

Institutions: Broad, TIGR, VectorBase

Duration: 1 day

Activity: A conference call did occur to review the draft project plan and discuss how best to process edits and arrive at a ratified document.

Task: Review and Acceptance of Project Plan

Institutions: Broad, TIGR, VectorBase, NIAID

Duration: 2.4 weeks

Activity: Acceptance by the centers can occur informally by email or conference call discussion. Once the project plan is edited and ratified by the centers, it will be reviewed by NIAID and a final version distributed.

Stage 1: Annotation Preparation (4 weeks)

Relatively simple preparation for the Stage 3 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: Create Central Repository (CR)

Institution: TIGR

Duration: 2 weeks

Activity: TIGR's IT department will create an FTP directory with password protection. The purpose of this site will be to serve as a central repository for data that is exchanged between the Broad, TIGR and VectorBase.

Task: Release August Assembly

Institution: Broad

Duration: 5 days

Activity: The Broad will work to submit the August assembly to GenBank by October 7.

In addition, both MSCs will make the assembly data available through their respective web sites, including project descriptions and basic search capabilities.

Task: Release useful datasets

Institutions: TIGR, Broad, VectorBase

Duration: 5 days

Activity: During the course of this project it is anticipated that the centers will share pre-annotation data sets that will be useful to the annotation effort. This data will be considered work product and not released to the scientific community. The information will be stored in a central password protected FTP site. Readme files will accompany the data files.

TIGR has produced and will make available the following data sets:

- Repeat library of Aedes-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: Run and Evaluate Computes

Institution: Broad

Duration: 3 weeks

Activity: To support the selection and manual annotation the 10 Mb quality-control regions, the Broad will run a set of independent computational analyses including repeat finding, BLAST against the NCBI non-redundant database, HMMER against PFAM, and several gene prediction algorithms.

Stage 2: Production Annotation Gene Structure (8 weeks)

Prior to generating this project plan, preliminary analysis of the Aedes genome sequence was performed on a preliminary assembly version. The results from this analysis were loaded into an annotation database to allow us to perform comprehensive evaluation of that information.

The data was evaluated for gene coverage by aligned and assembled EST sequences, exon/intron structure, transposon and repeat content. We were also able to generate a gene prediction training set based on the assembled EST sequences and assess the performance of multiple gene finders, including the effect of repeat masking on gene finder output.

Several features of the Aedes genomes were confirmed in this preliminary analysis:

- The irregular intron distribution suggested by the EST data is striking, with 47% smaller than 100 bp and 17% larger than 10kb.
- The genome contains many transposon-related sequences. These can be present within the introns of protein-coding genes, in addition to intergenic regions. Protein coding genes are also found in the introns of other genes, as reported in *Drosophila*.
- The fragmented assembly led to a high percentage of truncated genes. These tend to confound gene finding programs as gene finders may fail to call genes missing start or stop codons, and exons missing in gaps tend to throw off reading frames and affect splice site prediction.

These preliminary results have shown that optimizing the pipeline for Aedes will be an R&D effort, requiring multiple iterations of parameterization/training and evaluation to arrive at the best possible gene set.

Thus, the major challenges facing us during Stage 3 are definition of correct gene boundaries given long introns lengths and intervening transposon sequences, dealing with partial genes caused by the fragmentation of the assembly and comprehensive repeat identification and screening.

The data generated by the three centers in this Stage will be deposited in the central repository for use in Stage 4: Evaluation of Data, and lead to an active exchange of methods and data between TIGR, the Broad Institute and Ensembl as pipeline configuration progresses.

Task: TIGR gene structure annotation

Institution: TIGR

Duration: 8 weeks

Activity: The following tasks pertain to TIGR's gene structure annotation.

Task: Run autopipeline

Institution: TIGR

Duration: 4 weeks

Activity: TIGR's annotation pipeline, Eukaryotic Genome Control (EGC) 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 and nematodes. It provides a wrapper 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:

- GlimmerHMM (Majoros et al. *Bioinformatics*. 2004:2878-9)
- 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)
- Twinscan (Korf et al. *Bioinformatics. Suppl 1*:S140-8) using *Anopheles* as an informant
- Jigsaw (Allen et al. *Genome Res.* 2004:142-8) dependent on other gene finders and similarity data

The basic pipeline for gene structure annotation will consist of the best-performing of the gene finders listed above as well as a set of similarity-based computes. The datasets will 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 will most likely be generated based on Evidence Modeler, which synthesizes protein and EST alignments with the data from 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: Submit TIGR 0.5 in Central Repository

Institution: TIGR

Duration: 1 day

Activity: The results of TIGR's annotation pipeline, particularly those data types established in Stage 1, will be deposited in the Central Repository FTP site.

Task: Iterative improvement of gene set

Institution: TIGR

Duration: 3 weeks

Activity: TIGR 0.5 will represent the raw output of the TIGR pipeline, a set of consensus gene predictions produced by a program called Evidence Modeler, which synthesizes protein and EST alignments with the data from the individual gene prediction programs.

Due to the complex nature of this genome project, we anticipate that we will need to process iterative improvements to this initial data set. We will identify missed and misannotated genes by checking 0.5 predictions against the protein alignments from *Anopheles* and *Drosophila*, screening introns larger than 10kb for the presence of possible protein coding genes, and by comparing the predictions to the aligned and assembled ESTs.

Task: Verify against *Anopheles* and *Drosophila*

Institution: TIGR

Duration: 5 days

Activity: All significant *Anopheles* and *Drosophila* protein matches identified by AAT searches will be rerun with Genewise to produce full-length or partial gene model predictions. These Genewise results will be compared to our predicted gene set to identify intergenic alignments, suggesting missed genes, and one to many mappings, suggesting inappropriately split genes.

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

Task: Locate genes in introns of others

Institution: TIGR

Duration: 5 days

Activity: Most gene prediction software will not correctly predict nested and overlapping genes. During the pre-annotation analysis, many examples of protein coding genes in the introns of other protein coding genes were identified. Therefore, we will extract all intron sequences greater than 10kb, masked for repeats, and process separately to capture gene predictions that may have been missed.

Task: EST Data Incorporation

Institution: TIGR

Duration: 5 days

Activity: Once we are confident that the gene set is substantively complete, the PASA program will be used to compare the gene set with the aligned and assembled EST sequences. PASA will report any conflicts between EST alignments and the gene structures, and can process a number of automated updates to resolve those differences.

PASA also reports intergenic EST alignments. These will 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: Quality Control

Institution: TIGR

Duration: 5 days

Activity: 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 overprediction 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: Submit 0.5.1 in Central Repository

Institution: TIGR

Duration: 1 day

Activity: Subsequent improvements to TIGR's gene set will be deposited in Central Repository FTP site.

Task: VectorBase: gene structure annotation

Institution: VectorBase

Duration: 8 weeks

Activity: 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 also *Anopheles gambiae*.

VectorBase will make preliminary assembly data available to community. To this end, an initial Release 0.5 of *Aedes* annotation will be made available during this Stage. This annotation will be based on one or more of the initially trained gene finders. This data will be documented carefully to indicate to the scientific community that it is preliminary and will potentially have problems with gene boundaries, resolving gene structures in repetitive regions, and capturing partial gene models.

These gene predictions will not be submitted into Genbank and will not be tracked to Release 1.0.

Task: autopipeline

Institution: VectorBase

Duration: 4 weeks

Activity: Briefly Ensembl annotates from evidence, either cDNAs or ESTs from the organism itself or protein sequences from related organisms. There are three main sources genes:

- a. cDNAs that encode proteins (species-specific, or closely related species). These are placed on the genome using a fast matching tool (pmatch) followed by an accurate Genewise gene prediction parameterized to favor global (Met-to-STOP) matches.
- b. cDNAs where the protein sequence is not clear (e.g., due to the fact it is fragmented) and ESTs. Ensembl places the sequences using the fast and accurate matching tool exonerate (Slater and Birney, BMC Bioinformatics, 15, 6(1):31) and then clusters the resulting partial fragments which overlap to form transcript structures.

c. Regions of the genome where protein sequence from other organisms matches. An optimized BLASTX system is used to find these regions, and then Genewise again is used to predict genes.

These three sources of genes are then merged into one final set, with a priority information of (a) > (b) > (c). In the Aedes build we expect to experiment with the following new techniques to improve the build quality.

a. Met extensions. We have extended the Genewise model to aggressively find starting Mets wherever possible.

b. Phase and intron position. Phase and intronic position are usually highly conserved in orthologous genes, and this provides an orthogonal set of information to find protein coding genes.

c. Use of joint *Aedes/Anopheles* gene prediction. We expect that using the Aedes genome we will be able to jointly predict these genes in both organisms. We will use the SLAM tool from the Patcher group and also investigate new protein-HMM based methods in this area.

Task: make v 0.5 available via web

Institution: VectorBase

Duration: 4 weeks

Activity: VectorBase will make preliminary assembly data available to community for BLAST, SSAHA access.

The current VectorBase and Ensembl sites for Anopheles provides some sense of what the user can expect. The main information on genes will be displayed as for *Anopheles*. This includes a full, easily navigable 'genome browser' display, showing transcripts on chromosomes in the context of other features. There are also pages for genes, transcripts and proteins showing graphics, sequences (including options for various kinds of mark-up and simple downloading) and cross-references to other databases. Protein pages include graphical depictions of protein domains with links to appropriate databases.

Orthologous links between Anopheles and Aedes will be shown, and the Ensembl site will also display family clusters of proteins, and orthologs in a range of vertebrate and model animal species. If SNPs are called on Aedes, either in the assembly process or from ESTs (or both), we will both display the SNPs in VectorBase and Ensembl and also submit the SNPs to dbSNP. We can also show SNP locations on coding sequences and effects on proteins.

Users will be able to enter the site by starting with a genomic region, by text searching, or by BLAST or SSAHA searches of the genome and of predicted transcript and protein sets.

In addition to simple sequence download of specific genomic regions or gene sequences, the powerful BioMart tool will be available to formulate genome-wide queries and return sets of annotation or sequences.

Task: Submit 0.5 or 0.5+ into Central Repository

Institution: VectorBase

Duration: 1 day

Activity: The results of the VectorBase annotation pipeline, particularly those data types established in Stage 1, will be deposited in Central Repository FTP site.

At the time of constructing this project plan it was not clear if Version 0.5 will be the data that will be delivered to TIGR for Stage 4: Evaluation of Data.

Task: Broad: 10 Mb annotation

Institution: Broad

Duration: 8 weeks

Activity: Rather than run an automated gene prediction pipeline, the Broad Institute will focus its effort on performing manual annotation on the 0.5 release over a selected region of the Aedes genome (approximately 10 Mb). This will provide a complementary data set that can be used to assess the strengths and weaknesses of the automated pipelines. The Broad Institute will also assist in determining the best methods of evaluating pipeline outputs and producing a robust, unified data set.

Task: Define and Annotate 10 Mb region

Institution: Broad

Duration: 8 weeks

Activity: A 5 Mb region of the Aedes genome to be annotated by Broad will be selected by our collaborators. The remaining 5 Mb region will be selected by the Broad annotation group based on the gene content and level of evidence used for annotating the selected region .

Task: Submit 10 Mb annotation in Central Repository

Institution: Broad

Duration: 1 day

Activity: The results of the Broad Institute's annotation pipeline, particularly those data types established in Stage 1, will be deposited in Central Repository FTP site.

Stage 3: Evaluation of Data (4 weeks)

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 will be deposited into Genbank in Stage 5.

Towards that end, there will be at least one round of prediction and comparison before a full set is provided. As with other genome projects of this size and scope, the final set might consist of the output of one center's pipeline, or by merging data from multiple data sets.

We also expect this approach will ensure that the annotation pipeline at TIGR is refined to operate at the highest possible quality, which will have a beneficial impact for all subsequent large genome projects.

There is also the possibility of incorporating a limited amount of manual curation submitted by the Aedes research community. This from VectorBase:

Manual annotation of the *Anopheles gambiae* genome currently follows two paths.

1. The Harvard group, under the direction of Bill Gelbart (FlyBase PI and VectorBase co-PI), manually annotates regions of the genome using the Apollo tool. The EnsEMBL group receives manual annotation of regions from the Harvard group as GAME-XML files. Gene structures are extracted and transferred to an EnsEMBL data schema (Gene, Transcript & Exon). The EnsEMBL group adds the predicted cDNA as supporting evidence for the prediction. These predictions are viewed as being 'Blessed' within the gene build and take precedence over all other gene models generated during the gene build.

2. Some additional manual annotation is submitted by members of the mosquito research community following a 'gene naming protocol' described on both the VectorBase and EnsEMBL Mosquito Genome sites. Basically, scientists work with genes of interest to determine potential gene structures and functions. These scientists then name genes based on their analysis, following the published naming protocol. These names, a virtual cDNA, and a summary of supporting evidence are submitted to EnsEMBL/VectorBase, where they are either approved, rejected, or sent back for suggested changes.

We envision that the Aedes community would follow a procedure similar to number 2 above. The idea would be that members of the Aedes community would provide potential gene structures (as virtual cDNAs) to VectorBase. We would map these to the assembly and extract a genomic region surrounding the gene for appraisal by Kathy Campbell, a VectorBase annotation scientist at Harvard. Approved regions and genes would be sent back to VectorBase as XML files and transformed into an EnsEMBL schema. Things to remember about this system include the fact that EnsEMBL stores coordinates of the predictions relative to a genome slice (and hence they are readily transferable between assemblies), with appropriate checks for the underlying sequence chunk. The EnsEMBL group have experience of all of these steps as this is essentially the system in place for dealing with the Anopheles data.

EnsEMBL/VectorBase would be responsible for all information coming through this path, with Aedes biologist Dave Severson providing review. These annotated genes would then be a part of the EnsEMBL/VectorBase contribution to the final merge.

We do not anticipate that a large amount of annotation would enter the Aedes annotation version 1.0 through this path, but it would be unfortunate to simply ignore good annotation data if they are available before the merge.

Task: implement work in proposal for evaluating and comparing gene sets

Institution: TIGR

Duration: 3 weeks

Activity: The tasks required for data evaluation include:

- Calculate sensitivity and specificity statistics vs. EST alignments and derived data (apparent full length genes).
- Calculate sensitivity and specificity statistics vs. manually curated 10 MB region.
- Perform direct comparisons between gene sets, enumerating identical matches, overlapping but different gene calls, and unique gene calls in each gene set.
- Generate a set of conserved genomic regions between Aedes-Anopheles and Aedes-Drosophila using a BLAST or MUMmer-based method. Assess how many of these conserved regions coincide with exon calls.
- Report results to centers towards developing a merging strategy, if one is required.

Task: define merge strategy based on evaluation

Institutions: TIGR, Broad, VectorBase

Duration: 5 days

Activity: Centers will review the results of data evaluation with input from the Aedes community and come to a decision on the composition of the final gene set.

Task: Possibility of Holiday Related Delay

Institutions: TIGR, Broad, VectorBase

Duration: 1 day

Activity: This task is meant to serve as a placeholder. If this schedule holds, data evaluation is scheduled to take place over the Winter Solstice/Christmas/New Year holiday period. Since we do not know how many critical personnel will be communing with snow covered forests or making obligatory family visits, communication between centers may be difficult and cause a delay. The consortium may wish to identify the exact numbers of hours used for the purpose of secular and non-secular celebration prior to entering into this task and the remaining project plan timeline will be adjust accordingly.

Stage 4: Data Generation Gene Structure v1.0 (4 weeks)

Once the best combined strategy is defined, we will generate the required data, 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 through TIGR.

Task: Implement Merging Strategy

Institution: TIGR

Duration: 3.6 weeks

Activity: Once the strategy for developing a single, robust gene set for Genbank release is agreed upon, TIGR will implement the strategy, leveraging existing tools for the automated update of gene models.

This dataset will be re-evaluated against the original measures of success, if substantial changes were made, and distributed to all centers for review.

Task: UTR Addition based on EST data

Institution: TIGR

Duration: 2 days

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

Task: Possibility of Annotation Tracking to New Assembly

Institution: TIGR

Duration: 1 day

Activity: Placeholder: If an improved assembly were to become available during the period reflected in this annotation project plan, additional time would be required to track the annotation forward to the new assembly. If this were to occur, we would assign a time period to this task and the rest of the timelines should adjust accordingly.

There is no formal plan to update the assembly, but there is the possibility that new or existing data will suggest ways that the assembly could be improved. Any further work on the assembly would need to be discussed between centers and approved by NIAID.

Stage 5: Production Annotation Functional Computes (3 weeks)

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

Task: Automated Gene Product Name Assignment

Institution: TIGR

Duration: 3 weeks

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

- 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

Gene product names will 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: Automated Gene Ontology Assignments

Institution: TIGR

Duration: 3 weeks

Activity: Gene products will also be assigned to Gene Ontology (GO) terms by transferring the curated GO associations of the *Drosophila melanogaster* ortholog and/or by using the mappings between PFAM/InterPro and gene ontology terms.

Stage 6: Genbank Submission v1.0 (2 weeks)

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 and will be maintained as accessioned objects in all subsequent releases of the *Aedes* genome. We anticipate that this approach will result in the best achievable set of genomic information for the user community that will be propagated into the public archives and maintained at VectorBase.

Task: Quality Control

Institution: TIGR

Duration: 5 days

Activity: 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, as outlined in Stage 3 above, we will also check gene name and Gene Ontology information for data and formatting problems.

Task: File Generation

Institution: TIGR

Duration: 5 days

Activity: The data for Release 1.0 is prepared for submission into Genbank. The data conveyed will be those data types that were described in Stage 1 and will most likely include: protein coding genes + UTRs, gene product names, and automated Gene Ontology assignments. We will 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.

Task: Genbank Submission 1.0

Institution: TIGR

Duration: 1 day

Activity: 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.

The primary contacts for this task will be Jennifer Wortman (jwortman@tigr.org) and Vish Nene (nene@tigr.org) at TIGR and, we anticipate, Karen Clark (kclark@ncbi.nlm.nih.gov) at NCBI.

Stage 7: Genbank Processing to Release (4 weeks)

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: Validated Genbank files sent to VectorBase for Processing

Institutions: TIGR

Duration: 1 day

Activity: 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 will be sent to VectorBase for processing so that the Genbank and VectorBase releases to the public can be coordinated.

Task: Creation and Coordination of Press Materials

Institutions: TIGR, Broad, VectorBase, NIAID

Duration: 4 weeks

Activity: Press releases should be prepared by the NIAID and the sequencing centers during the time that the Genbank record is processed by NCBI staff. Statistics of the Aedes genome will be readily available. This task will require coordination between the sequencing centers and the NIAID for timing of the announcement, points of contact, access to the data and consistency among the documents.

Task: Data Release/Press Release

Institutions: Broad, TIGR, VectorBase, NIAID

Duration: 1 day

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

Stage 8: Manuscript Preparation (29 weeks)

Broad, TIGR and VectorBase will participate in preparation of a scientific publication in a peer-reviewed journal. Note that this time period begins prior to the 1.0 Genbank release, and extends past the release date by 4-5 months.

Appropriate personnel will be utilized for production of analysis used for manuscript preparation during this Stage, and for the purpose of supporting scientific collaborators involved in manuscript preparation.

Stage 9 will also require extensive coordination among the centers, and the scientists participating in the scientific publication, represented by Dr. Severson. These activities include preparation of data, data exchange, interfaces creating access for collaborators, genome analysis by specialized working groups for analysis, submission of on-line information to the publisher and preparation of text.

Task: Define Publication Strategy

Institutions: TIGR, Broad, VectorBase

Duration: 5 weeks

Activity: It is recommended that an overall publication strategy be established for analysis of the Aedes genome.

This strategy will require the coordination of the following activities:

- Preparation of genome statistics
- Genome family analysis, domains, motifs
- Comparative analysis
- Metabolic pathway analysis
- QTL mapping associations with assembly
- Coordination of on-line information with peer-reviewed journal
- Writing and submission of manuscript into peer-reviewed journal

Dave Severson has spoken to journal editors and community members and is working with the centers to develop a comprehensive plan.

Task: Data Analysis for Publication

Institutions: TIGR, Broad, VectorBase

Duration: 20 weeks

Activity: A specific project plan for this time period will need to be developed based on the publication strategy and specific analyses required. This work may require targeted assembly improvements, integration of externally derived data and additional computational and manual analyses of the annotation data.

Task: Support for Collaborator Analysis

Institutions: TIGR, Broad, VectorBase

Duration: 20 weeks

Activity: The Broad, TIGR and VectorBase have strong track records of providing scientific support for collaborators participating in genome publications and of exchanging data with other large genome sequencing and annotation centers.

For example, TIGR has a software support system which allows users to log into password protected web pages giving access to genes, functional assignments, chromosomal elements, and physical clone information. This software has enabled researchers to perform analyses required for genome publications, as well as contribute data such as functional assignments and gene names. We expect the collaborative nature of open source development and the provision of thorough documentation will promote use of our data exchange technology.

Task: Write Paper

Institution:

Duration: 16 weeks

Activity: