

VectorBase: A Bioinformatics Resource Center for Invertebrate Vectors of Human Pathogens

Frank H. Collins¹ Robert Bruggner¹, E.O. Stinson¹, Pantelis Topalis², and Christos Louis²

1. Center for Tropical Disease Research and Training, University of Notre Dame, Notre Dame, IN 46556.
2. Institute of Molecular Biology and Biotechnology, FORTH, Crete

VectorBase is a centralized relational database and web interface focused on invertebrate vectors of human pathogens.

The GMOD database schema, Chado, has been chosen to be used for **VectorBase's** core schema. Chado, originally conceived as the next generation **FlyBase** database, will be used to store all of **VectorBase's** genomic data as well as the accompanying metadata such as annotation information, gene expression information, etc. The relational database management system (RDBMS) PostgreSQL is platform on which the Chado schema has been implemented. As public domain software, PostgreSQL offers extensive functionality and stability without issues of costly licenses for commercial RDBMS engines.

VectorBase will be guided largely by features and goals embraced by two of its core partners, **FlyBase** and Ensembl as well as embracing the design goals of the open source Genome Model Organism Database Construction set, GMOD (<http://www.gmod.org/>).

VectorBase will manage, display, and analyze data for all vectors for which genome level data sets are developed (genome sequences, extensive EST sequence sets, other large scale genome-derived data sets, or data sets based on functional analysis of the genome). Browsing of **VectorBase's** genomic data will be via the Ensembl Genome Browser, made possible by a new Ensembl database adaptor designed exclusively for the Chado schema. Other key data store and analysis components of **VectorBase** will be provided by the European Molecular Biology Laboratory in Heidelberg, Germany, the Institute of Molecular Biology and Biotechnology (IMBB) in Heraklion, Crete, the **FlyBase** group at Harvard University, and the Center for Tropical Disease Research and Training at the University of Notre Dame. These institutions will also represent a major conduit to the scientific community that will facilitate community use of **VectorBase**. They will also help develop the level of coordinated scientific interest in a particular vector that could lead to a genome project. In addition, **VectorBase** will recognize these consortia as primary sources of formal guidance.

During the first 5 year cycle, **VectorBase** will provide a bioinformatics resource for the display and analysis of all genome data that relate to arthropod vectors of human pathogens. Our current expectation is that this will include as a minimum data related to complete draft genome sequences for the following

species: *Anopheles gambiae*, *Aedes aegypti*, *Culex pipiens quinquefasciatus*, *Ixodes scapularis*, and *Glossina morsitans*. We also expect to manage lower coverage draft sequences (3-5 X) for a number of other *Anopheles* species in the *Anopheles gambiae* complex as well as potentially some additional vector genomes.