

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

Frank H. Collins, Center for Tropical Disease Research and Training, University of Notre Dame, Notre Dame, IN 46556.

VectorBase will be a centralized relational database based on a publicly available and supported database management system with a Structured Query Language (SQL) that can be searched using query by example (QBE). All information records in **VectorBase** will use XML or another explicitly defined, parsable format that includes a process for data representation (i.e. Document Type Definition - DTD).

VectorBase will use an open-source relational database management schema like MySQL or PostgreSQL. It will be based in large measure on relational database features and goals embraced by two of its core partners, FlyBase and the Ensembl Mosquito Genome Browser. **VectorBase** will embrace the Genome Model Organism Database Construction set (GMOD) design goals (<http://www.gmod.org/>). The GMOD schema, Chado, is being considered as the core schema for **VectorBase** for the biological annotation aspect of information, in common with FlyBase. Chado is a modular schema for handling biological data intended to be used as both a primary datastore schema as well as a warehouse style schema. Chado was originally conceived as the next generation FlyBase database, combining the sequence annotation database Gadfly with the Harvard and Cambridge components of FlyBase. The genome sequence portion of **VectorBase** will probably use an Ensembl-based schema, which is also affiliated to GMOD. The Ensembl genome schema can be used as a modular component of Chado. As public domain software, MySQL and PostgreSQL offers the advantages of being mirrored at many sites without issues of costly licenses for commercial RDBMS engines.

Data and their associated analyses and annotation will be entered into **VectorBase** through a distributed set of data management tools. While the exact nature of these tools has not yet been determined, they will probably be organized along both functional and taxonomic (different vector species) lines. The GMOD distributed annotation system (DAS) will be the client-server system through which the single client **VectorBase** will integrate information from multiple servers. The Ensembl system is already a compliant DAS server and client. The DAS system will allow **VectorBase** to gather up genome and genome-related information from multiple distant web sites, collate the information, and display it to the **VectorBase** user in a single view. The distributed information providers will use data systems designed to be replicated at different sites where individual, specialized research groups can focus on the different organisms whose data will populate **VectorBase**. We anticipate several different types of distributed information providers based on pre-existing models, as described in detail below. All contributors to **VectorBase** are committed to the GMOD plan for shared, modular database structures, and all tools developed *de novo* or significantly re-designed for **VectorBase** will follow GMOD guidance.

The central feature of **VectorBase** will be based on the Ensembl genome display and analysis tool developed and managed by the European Bioinformatics Institute (EBI). Other key 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, a population biology laboratory at UCLA, and the Center for Tropical Disease Research and Training at the University of Notre Dame. **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). Moreover, **VectorBase** will assume responsibilities for developing and updating internationally recognized Reference Data Sets for the organisms included, such as the Reference Sequence sets that comprise the official copies of specific genomes as defined and maintained by NCBI.

VectorBase will initially focus on five vector species for which genome projects are in process or in development, but it will be designed to rapidly expand to include new vector genomes as such projects are initiated and resources identified. The design of **VectorBase** has been modeled in many respects after FlyBase, which is a lead participant in the GMOD project. **VectorBase** will take an active role in helping to establish formal international consortia of scientists who represent particular vector species or groups of species. Such consortia will serve several roles. They will 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. The present model for these consortia is the International *Anopheles gambiae* Sequence Committee that was established to manage issues related to the *An. gambiae* genome sequencing project.

VectorBase is a community service tool that is currently under development. Queries and suggestions should be sent to Frank H. Collins (Frank@nd.edu).