*Published online 21 November 2008* *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D583–D587***
*doi:10.1093/nar/gkn857*

# VectorBase: a data resource for invertebrate vector genomics

**Daniel Lawson[1,*], Peter Arensburger[2], Peter Atkinson[2], Nora J. Besansky[3], Robert V. Bruggner[3], Ryan Butler[3], Kathryn S. Campbell[4], George K. Christophides[5], Scott Christley[3], Emmanuel Dialynas[6], Martin Hammond[1], Catherine A. Hill[7], Nathan Konopinski[3], Neil F. Lobo[3], Robert M. MacCallum[5], Greg Madey[3], Karine Megy[1], Jason Meyer[7], Seth Redmond[5], David W. Severson[3], Eric O. Stinson[3], Pantelis Topalis[6], Ewan Birney[1], William M. Gelbart[4], Fotis C. Kafatos[5], Christos Louis[6,8] and Frank H. Collins[3]**

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [2]Department of Entomology, University of California, Riverside, 900 University Avenue, Riverside, CA. 92521, [3]Center for Global Health and Infectious Diseases, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46656-0369, [4]The Biological Laboratories, 16 Divinity Avenue, Harvard University, Cambridge, MA 02138, USA, [5]Cell and Molecular Biology Department, Imperial College London, South Kensington Campus, London SW7 2AZ, UK, [6]Institute of Molecular Biology and Biotechnology, FORTH, Vassilika Vouton, PO BOX 1385, Heraklion, Crete, Greece, [7]Department of Entomology, Purdue University, West Lafayette, IN 47907, USA and [8]Department of Biology, University of Crete, Heraklion, Crete, Greece

Received September 15, 2008; Revised and Accepted October 16, 2008

## ABSTRACT

**VectorBase (http://www.vectorbase.org) is an NIAID-funded Bioinformatic Resource Center focused on invertebrate vectors of human pathogens. VectorBase annotates and curates vector genomes providing a web accessible integrated resource for the research community. Currently, VectorBase contains genome information for three mosquito species: *Aedes aegypti*, *Anopheles gambiae* and *Culex quinquefasciatus*, a body louse *Pediculus humanus* and a tick species *Ixodes scapularis*. Since our last report VectorBase has initiated a community annotation system, a microarray and gene expression repository and controlled vocabularies for anatomy and insecticide resistance. We have continued to develop both the software infrastructure and tools for interrogating the stored data.**

## INTRODUCTION

VectorBase is a genome information system which provides a genome browser for visualizing genome annotations, including DNA and protein alignments, variations, protein feature data and functional data sets, such as microarray expression analysis. We are active in producing genome annotation ourselves but also collaborate with a range of partners including our sister Bioinformatic Resource Center's (1) to incorporate and improve the annotations.

The reduction in cost of sequencing has seen genomes become available for an increasing number of vector species. VectorBase is directly responsible for three mosquito species (*Aedes aegypti*, *Anopheles gambiae* and *Culex quinquefasciatus*) and the tick *Ixodes scapularis*. We work closely with the genome sequencing centres on the initial annotation and publication of these genomes and then assume responsibility for ongoing re-annotation tasks. A number of other genomes are within scope for VectorBase including the body louse (*Pediculus humanus*), triatomine bug (*Rhodnius prolixus*), tsetse fly (*Glossina morsitans morsitans*) and sand flies (*Lutzomyia longipalpis* and *Phlebotomus papatasi*). A full list of VectorBase species and data sets can be accessed on the website (http://www.vectorbase.org/Help/Current_release).

This report highlights the new genomes integrated into VectorBase and some of the new features and improvements that we have added since our last report (2). Users interested in the VectorBase project should visit the main web page or help pages (http://www.vectorbase.org/Help/Main_Page) for more information about the project.

*To whom correspondence should be addressed. Tel: +44 1223 494 444; Fax: +44 1223 494 468; Email: lawson@ebi.ac.uk

## ACCESSING VECTORBASE

VectorBase as a web resource is linked with a number of other databases, most notably the public nucleotide and protein databases. Direct cross-references to the genes, transcripts and proteins exist in the GenBank/EMBL/DDBJ genome assembly records as well as the UniProt protein records, where both *An. gambiae* and *Ae. aegypti* are deemed to be complete proteomes. Other resources which use VectorBase data range from large general resources, such as Ensembl (http://www.ensembl.org) and Refseq (http://www.ncbi.nlm.nih.gov/RefSeq) to the more biologically focused proteinase database Merops (http://merops.sanger.ac.uk) and miRNA target predictions in mirBase (http://microrna.sanger.ac.uk). The VectorBase site and wiki resource are indexed by the major search engines allowing users to readily find content of interest.

## EXPANDED ROLE OF VECTORBASE

VectorBase is active in all stages of genome analysis including initial annotation of new genome sequences in collaboration with the sequencing centres, such as JCVI and The Broad Institute and subsequent re-annotation using both computational and manual approaches in liaison with the community. Automated annotation using the Ensembl system (3) was undertaken for the new genomes (*C. quinquefasciatus*, *P. humanus* and *I. scapularis*). The process of resolving differences between VectorBase and the partner sequencing centre annotations has been a fruitful task leading to high-quality automated annotation but problems will remain which can only be addressed using further resources (expressed sequence tags or new genome sequences) or through manual appraisal of the automated gene predictions. VectorBase has invested some resource toward the latter and implemented strategies for involving the community in the annotation effort. We have also implemented data mining tools, such as the HMMER package (http://hmmer.janelia.org/) to build profile hidden Markov models from multiple sequence alignments which can then be used for sensitive database searching using statistical descriptions of a sequence families consensus.

## EXTENSIVE MANUAL APPRAISAL OF MOSQUITO GENE MODELS

The annotation of the *An. gambiae* genome is being manually appraised using the GMOD annotation tool Apollo (4). Currently, over 50% of the genome has been completed including the entirety of the chromosome arms 2L, 2R and X. Many loci have been updated to correct systematic errors in the computational annotation; especially in reference to tandem arrays of multi-gene families, gene merges from multiple partial predictions and the removal of suspect predictions likely to be based on transposable element sequences. Manual annotations are stored in a separate CHADO database (5), displayed as a track in the genome browser via DAS (6) and integrated into the main gene build during the next round of re-annotation.

Small-scale manual appraisal of gene predictions has been undertaken for *An. aegypti* and *C. quinquefasciatus* as part of the quality control for the gene builds. In the case of *C. quinquefasciatus*, this revealed at least 1500 predictions which were removed from the CpipJ1.2 dataset. Amongst the deprecated gene predictions were a large set of single exon predictions which had no supporting transcript evidence and no similarity to other mosquito proteomes or any other sequences in the public databases. Expert opinion was that these were erroneous over-prediction by the computation algorithms rather than a large Culex-specific gene family. Efforts such as these highlight our determination to improve gene prediction accuracy through the integration of new data sets and the re-appraisal of the existing prediction set.

## COMMUNITY ANNOTATIONS

VectorBase employs community representatives focused around the NIAID-funded species (the three mosquito genomes and *I. scapularis*). The representatives were hired from within the relevant community and have both biological knowledge of the species and informatics skills. Their role is to liaise with the community providing helpdesk and training capacity, acting as mediators and quality assurance for data submission of gene predictions and as advocates for the user community in the development of the VectorBase resource.

We have developed a Community Annotation Pipeline (CAP) to facilitate community involvement in the curation of these genomes. This system consists of a CHADO database which stores annotations, both from the manual effort within VectorBase and those submitted directly from the community, and a web interface submission tool to upload data. Submitters use a spreadsheet format and can include gene predictions, gene symbols and gene descriptions, and attach GO terms or citations to a gene model. One aspect of the submission system is its ability to align a cDNA sequence to the genome using exonerate (7). The simplicity of the submission process in conjunction with community representative involvement in data quality consistency checks (e.g. does the submitted sequence translate correctly) ensures that any required discussion and error correction happens in a timely manner. Currently, the CAP system contains over 13 000 gene predictions. This system replaces the old Anopheles Symbol Database hosted by Ensembl and interested users can find more information about the CAP on the website (http://www.vectorbase.org/Help/Community_Annotation:Submission_User_Guide).

## COMPARATIVE ANALYSIS OF THE MOSQUITO GENOMES

The availability of the 'Culex' genome annotation facilitates comparison of the three main families of mosquitoes (Anopheline, Aedine and Culicine) with the model dipteran *Drosophila melanogaster*. As before VectorBase has

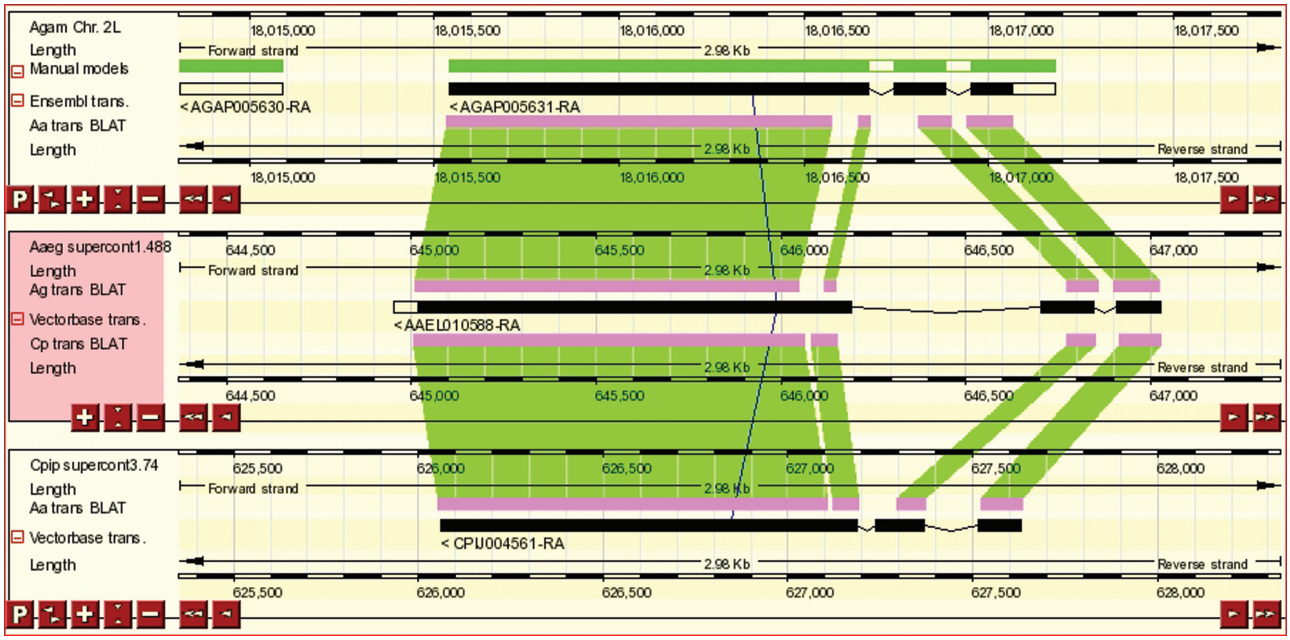


**Figure 1.** Three-way comparative display of the mosquito genomes *An. gambiae*, *Ae. aegypti* and *C. quinquefasciatus* (top panel to bottom panel, respectively). Gene predictions are coloured black with the regions of similarity between the genomes indicated by the pink boxes where similarity is indicated by the connection of blocks using the green lines. The upper panel, *An. gambiae*, also shows a manual annotation confirming the prediction and that the refinement of predictions in one species can be leveraged to improve the quality of prediction across all the mosquito genomes.

calculated pairwise tBLAT alignments (8) that can be used to connect between the genomes in a multi-contig view (Figure 1). Multi-contig views are available using the 'View alongside' option in the left-hand navigation panel and as links in the Gene Ortholog section of gene pages. We have added multiple genomic alignments calculated using Pecan (http://www.ebi.ac.uk/~bjp/pecan), which has been shown to be one of the best algorithms in terms of specificity and sensitivity (9). For each genomic position, the level of evolutionary constraint has been evaluated using GERP (10) and stretches of Pecan alignment showing a high level of conservation are marked as constrained elements.

Orthologs and paralogs are calculated using Ensembl Compara GeneTrees pipeline. This method is based on maximum likelihood phylogenetic trees built by TreeBest (http://treesoft.sourceforge.net). The trees, presenting the evolutionary history between the genes, are reconciled with the species trees and help in differentiating between duplication and speciation events. The gene tree for a particular gene accessible via the left-hand navigation menu of the gene page and ortholog/paralog data are available for querying via the BioMart interface (11).

## GENE EXPRESSION DATABASE

VectorBase has continued to develop its microarray experiment repository and gene expression reports (http://funcgen.vectorbase.org/ExpressionData/). The database now contains 12 array designs (including Affymetrix) and 17 experiments. We continue to actively solicit the community for microarray data which is reflected by the fact that half of the available experiments were published last year. The repository is built around the BASE database system (http://base.thep.lu.se/) and is integrated with the genome via probe mappings in the browser and navigation links for each gene in the GeneView pages (Figure 2). Other new features include text-based search, publication-quality plots and a more robust and extendible implementation of the statistical analysis.

## ONTOLOGIES AND CONTROLLED VOCABULARIES

VectorBase has developed controlled vocabularies describing the anatomy of mosquitoes (Taxonomist's Glossary of Mosquito Anatomy, TGMA), and ticks (Tick Anatomy by Dan Sonenshine, TADS) (12). These ontologies are fully compliant with CARO, the common reference ontology for anatomy (13) and contain 1861 and 628 terms, respectively. The ontologies can be browsed through the website (http://www.vectorbase.org/Search/CVSearch/) allowing for the concurrent visualization of the anatomical structures described. The ontologies are also available from the Open Biomedical Ontologies (OBO) Foundry (http://www.obofoundry.org) which acts as a central repository for all science-based ontologies. The two ontologies enhance the possibilities for annotation of gene expression experiments performed on disease vectors (14).

### MIRO and IRBase

AnoBase, the precursor database of VectorBase, has been making available data on insecticide resistance for a number of years (15). VectorBase has expanded this

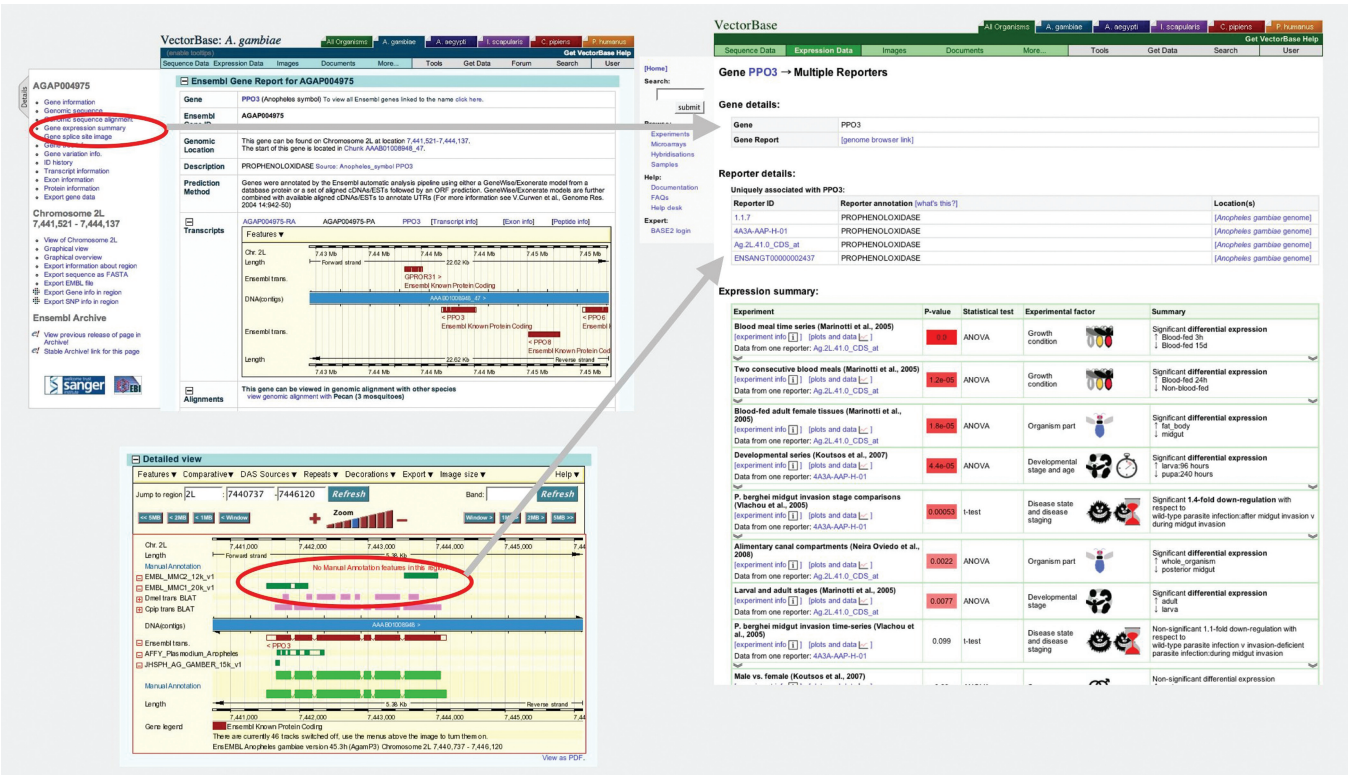


**Figure 2.** Integration of the genome browser with the microarray gene expression database. Links from gene pages, highlighted in red in the top left navigation panel, connect to a summary of the gene expression in all available experiments (right-hand panel). Similarly, there are direct links from the mapped reporter probes on the genome browser (bottom left panel) to the gene expression summary based on inferred overlap of the reporter and the current genome annotation. The gene expression summary lists experiments for which the gene has data ranked by statistical significance with a brief summary of the experimental details and links to more detailed analysis.

resource, building an ontology (MIRO) that describes features associated with insecticide resistance. This ontology was then used to upgrade the relevant database section with an enhanced search capability. Named IRBase, this can now be searched at http://anobase.vectorbase.org/ir/. Additional data based on new studies as well as on existing published ones are currently being integrated into IRBase, and this tool will be developed into the global database on insecticide resistance for disease vectors.

## OUTREACH

VectorBase staff regularly attends teaching workshops to demonstrate the database and the tools available for browsing and query the data. Recent workshops locations include Brazil, Kenya, Mali and South Africa. As the VectorBase genome browser is powered by the Ensembl system their extensive outreach program is also applicable to our databases, potentially giving access to numerous training courses worldwide throughout the year. VectorBase is continuously developing outreach and documentation resources which include the help contact e-mail (info@vectorbase.org) as well as a quarterly newsletter, a FAQ (frequently asked questions), Help Wiki (http://www.vectorbase.org/Help/Main_Page) and a community forum (http://www.vectorbase.org/sections/Forum/index.php). More details relating to these and other resources can be found on the website.

## FUTURE DEVELOPMENTS

Massively parallel sequencing technologies (both pyrosequencing and sequencing by synthesis) are being adopted by the sequencing centres reducing the cost of genome sequencing. We expect that this will speed up the generation of genome sequences from vector species which hitherto have been low priority because of size, cost or the practicalities of DNA availability. A number of Anopheles species will be targeted for genome sequencing (http://www.vectorbase.org/Docs/ShowDoc/?doc=White Papers) and the reduction in cost means that individual labs can produce significant amounts of sequence data from species or isolates. The integration and management of these data will be a major challenge for the coming years.

Analysis of populations and variation will increase at the sequence level and VectorBase will continue its partnership with Ensembl to process, store and represent this data. Population studies involve other types of data (including insecticide resistance, epidemiology, environmental conditions and vectorial transmission), which are not currently part of the VectorBase data schema and so we will work with partners in the relevant fields to

integrate these data with VectorBase enhancing the utility of the resource to the vector genomics community.

## ACKNOWLEDGEMENTS

We acknowledge the many researchers that have provided data through the Community Annotation Pipeline and thank the reviewers for helpful discussions.

## FUNDING

NIAID (contract HHSN266200400039C to the core VectorBase project); BioMalPar network of excellence (to the core VectorBase project). Funding for the open access charge: NIAID.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Greene,J.M., Collins,F., Lefkowitz,E.J., Roos,D., Scheuermann,R.H., Sobral,B., Stevens,R., White,O. and Di Francesco,V. (2007) National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. *Infec. Immun.*, **75**, 3212–3219.
2. Lawson,D., Arensburger,P., Atkinson,P., Besansky,N.J., Bruggner,R.V., Butler,R., Campbell,K.S., Christophides,G.K., Christley,S, Dialynas,E. *et al.* (2006) VectorBase: a home for invertebrate vectors of human pathogens. *Nucleic Acids Res.*, **35**, D503–D505.
3. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D514.
4. Lewis,S.E., Searle,S.M.J., Harris,N., Gibson,M., Iyer,V., Ricter,J., Wiel,C., Baytraktaroglu,L., Birney,E., Crosby,M.A. *et al.* (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, R82.
5. Zhou,P., Emmert,D. and Zhang,P. (2006) Using Chado to store genome annotation data. *Curr. Protoc. Bioinformatics*, Chap 9, Unit 9.
6. Jenkinson,A.M., Albrecht,M., Birney,E., Blankenburg,H., Down,T., Finn,R.D., Hermjakob,H., Hubbard,T.J., Jimenez,R.C., Jones,P. *et al.* (2008) Integrating biological data - the distributed annotation system. *BMC Bioinformatics*, **22**, S3.
7. Slater,GS and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
8. Kent,W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
9. Margulies,E.H., Cooper,G.M., Asimenos,G., Thomas,D.J., Dewey,C.N., Siepel,A., Birney,E., Keefe,D., Schwartz,A.S., Hou,M. *et al.* (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.*, **17**, 760–774.
10. Cooper, GM, Stone EA, Asimenos G., NISC Comparative Sequencing Program, Green ED, Batzoglou S, and Sidow A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
11. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
12. Topalis,P., Tzavlaki,C., Vestaki,K., Dialynas,E., Sonenshine,D.E., Butler,R., Bruggner,R.V., Stinson,E.O., Collins,F.H. and Louis,C. (2008) Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase. *Insect Mol. Biol.*, **17**, 87–89.
13. Haendel,M.A., Neuhaus,F., Osumi-Sutherland,D.S., Mabee,P.M., Mejino,J.L.V., Mungall,C.J. and Smith,B. (2008) CARO - the Common Anatomy Reference Ontology. In Burger,A., Davidson,D. and Baldock,R. (eds), *Anatomy Ontologies for Bioinformatics: Principles and Practice* vol. Chapter 16, Springer, New York, vol. Chapter 16, pp. 311–333.
14. Topalis,P., Lawson,D., Collins,F.H. and Louis,C. (2008) How can ontologies help vector biology? *Trends Parasitol.*, **24**, 249–252.
15. Topalis,P., Koutsos,A.C., Dialynas,M., Kiamos,C., Hemingway,J. and Louis,C. (2005) AnoBase: a genetic and biological database of anophelines. *Insect Mol. Biol.*, **14**, 591–597.