Published online 1 December 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D503–D505***
doi:10.1093/nar/gkl960

VectorBase: a home for invertebrate vectors of human pathogens

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

European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK, [1]Department of Entomology, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA, [2]Department of Biological Sciences, Center for Global Health and Infectious Diseases, University of Notre Dame, Notre Dame, IN 46656-0369, USA, [3]The Biological Laboratories, 16 Divinity Avenue, Harvard University, Cambridge, MA 02138, USA, [4]Department of Cell and Molecular Biology, Imperial College London, South Kensington Campus, London SW7 2AZ, UK, [5]Institute of Molecular Biology and Biotechnology, FORTH, Vassilika Vouton, PO Box 1385, Heraklion, Crete Greece, [6]Department of Entomology, Purdue University, West Lafayette, IN 47907, USA, [7]Department of Genetic Medicine and Development, Swiss Institute of Bioinformatics, University of Geneva Medical School, 1 rue Michel-Servet, 1211 Geneva, Switzerland and [8]Department of Biology, University of Crete, Heraklion, Crete, Greece

Received September 15, 2006; Revised October 23, 2006; Accepted October 24, 2006

ABSTRACT

VectorBase (http://www.vectorbase.org/) is a web-accessible data repository for information about invertebrate vectors of human pathogens. VectorBase annotates and maintains vector genomes providing an integrated resource for the research community. Currently, VectorBase contains genome information for two organisms: *Anopheles gambiae*, a vector for the *Plasmodium* protozoan agent causing malaria, and *Aedes aegypti*, a vector for the flaviviral agents causing Yellow fever and Dengue fever.

INTRODUCTION

Even before the completion of the human genome a number of laboratories initiated projects to sequence the genomes of important human pathogens: *Plasmodium*, *Trypanosome* and *Leishmania* species (1–3). The aim of these projects was to better understand the biology of the pathogen through its genome, with the goal of identifying new therapeutics and thus shorten the time from therapeutic lead to marketable product, a notoriously slow process. A more holistic approach to improving our understanding of these pathogens needs to include intermediary vectors where they exist. Over the past few years the cost of genome sequencing has fallen dramatically making it feasible to sequence the genomes of vectors and complete our knowledge of the triumvirate of species involved in many parasitic diseases.

VectorBase is funded by the National Institute of Allergy and Infectious disease (NIAID) as part of a group of Bioinformatics Resource Centres (BRCs) (http://www.brc-central.org/) aiming to provide web-based resources to the scientific community for organisms considered to be causing or transmitting emerging or re-emerging infectious disease. Parallel to this, NIAID has funded a number of genome projects of important vector species that are destined to be housed within the VectorBase system (Table 1).

VectorBase is involved in all the stages of genome analysis: first-pass annotation of new genome sequences in collaboration with the sequencers, re-annotation of existing genome sequences and submission of these data sets to the public nucleotide databanks.

VectorBase acts as the repository for the genome and predicted gene set providing web access for browsing and data mining capability. VectorBase participates in teaching workshops (supporters include WHO-TDR, MR4, EMBO and BioMalPar) and has undertaken 'hands-on' demonstrations at international meetings. VectorBase strives to improve

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

Table 1. List of vector species scheduled for inclusion into VectorBase

Vector	Disease	Status
Aedes aegypti[a]	Yellow and Dengue fever	Complete
Anopheles gambiae PEST	Malaria	Complete
A.gambiae M form[b]	Malaria	Initiated
A.gambiae S form[b]	Malaria	Initiated
Culex pipiens quinquefasciatus[a]	Lymphatic filariasis	Assembly
Glossina morsitans[c]	Sleeping sickness	Initiated
Ixodes scapularis[a]	Lyme disease	Sequencing
Lutzomyia longipalpis[d]	Leishmaniasis	Planned
Pediculus humanus[b]	Typhus	Initiated
Phlebotomus papatasi[d]	Leishmaniasis	Planned
Rhodnius prolixus[b]	Chagas disease	Initiated

[a]Funded by NIAID.
[b]Funded by NHGRI.
[c]Funded by Wellcome Trust.
[d]Funded by NHGRI and Wellcome Trust.

the accuracy and scope of the annotations, expanding controlled vocabularies for the vectors and incorporating new data types (expression, population and variation data).

RESULTS

Data storage

VectorBase uses the Genome Model Organism Database (GMOD) construction set for the storage of genome sequence and annotations. The GMOD CHADO schema facilitates the rapid incorporation of diverse data types, e.g. literature, controlled vocabularies and good inter-operability with the manual annotation effort, using Apollo, as well as the Ensembl system. For web display and access to the genome data and annotation, VectorBase utilises the Ensembl database schema, API and web code (4).

Genome browsing and data mining

The Ensembl system provides a good model for handling genomic data from a number of species in a consistent and unified manner and a highly sophisticated set of interlinked web pages.

Entry points into the genome are through text searches of gene names, symbols and descriptions, pairwise similarity searches and from cross-references in the public nucleotide and protein databanks.

The VectorBase website contains standard Ensembl style gene and transcript pages. Gene pages contain information about the prediction including gene orthologue and protein features (signal peptides, *trans*-membrane domains, InterPro domains). Gene Ontology (GO) codes and Enzyme Classification (EC) numbers are assigned where possible.

Batch file downloads are available for both the raw sequence data (fasta files of genomic sequence, including repeat masked sequence and ESTs) and the annotation (GFF3 files or a MySQL dump for use with the Ensembl API).

Batch searching capabilities are handled by the powerful data mining tool BioMart (5) and through two spreadsheets, AnoXcel and AegyXcel (6) that contain gene based information about the presence of signal peptides, *trans*-membrane domains, protein domains and the best similarity with yeast, *Drosophila* and human.

Annotation

The two genomes currently available through VectorBase are good examples of the multiple roles undertaken by the group.

Anopheles gambiae annotation

The *A.gambiae* PEST genome was published in 2002 (7) with a genome size of 260 Mb and ~14 000 genes. The predicted gene set has been reviewed and updated several times since publication. This process involves a blend of automated evidence-based gene prediction (8) and manual approaches. Manual appraisal of gene models is firstly targeted to regions of interest from the community and regions for which we are aware that automated approaches fail. A manual re-annotation of chromosome arm 2L is finished. Manually appraised gene models are highlighted for the user as a separate track on the genome browser.

Aedes aegypti annotation

The *A.aegypti* Liverpool strain was sequenced and assembled by The Institute of Genomic Research (TIGR) and the Broad Institute. *Aedes* has a large genome size of 1.3 Gb and a predicted gene complement of ~16 000 genes. This represents the first-pass annotation of the genome using automated approaches. Improvements in quality will be achieved by manual efforts and enhancements to the evidence-based automated gene predictions.

Sequence comparisons between *Anopheles* and *Aedes*

The presence of two related mosquito genomes in VectorBase allows for comparative analysis to identify conserved regions between the genomes. This can be useful in verifying, and correcting, gene models and for studying gene family expansions. Translated BLAT (9) similarities between the two mosquito species and with *Drosophila* have been identified. Figure 1 shows a view of an orthologous locus between the two mosquito genomes and highlights the expanded intron size in *A.aegypti* and related higher repeat content.

Use of Distributed Annotation System (DAS) in VectorBase

The DAS protocol (10) allows community researchers to integrate and display their data sets in the genome browser window. This is especially powerful with alternative sets of gene predictions. As an example, the *Anopheles* browser contains DAS tracks for alternate EST based gene predictions from AnoEST (11), an independent re-annotation effort by Li *et al*. (12) and predictions based on mass spectrometry data (13).

Microarray data

Microarray data exists for both *Anopheles* and *Aedes* and array probes from both species are mapped to the genome. These alignments are displayed in the browser and queries can be made against these via BioMart. Concise expression summaries for probes and genes are made available as experiments are published.

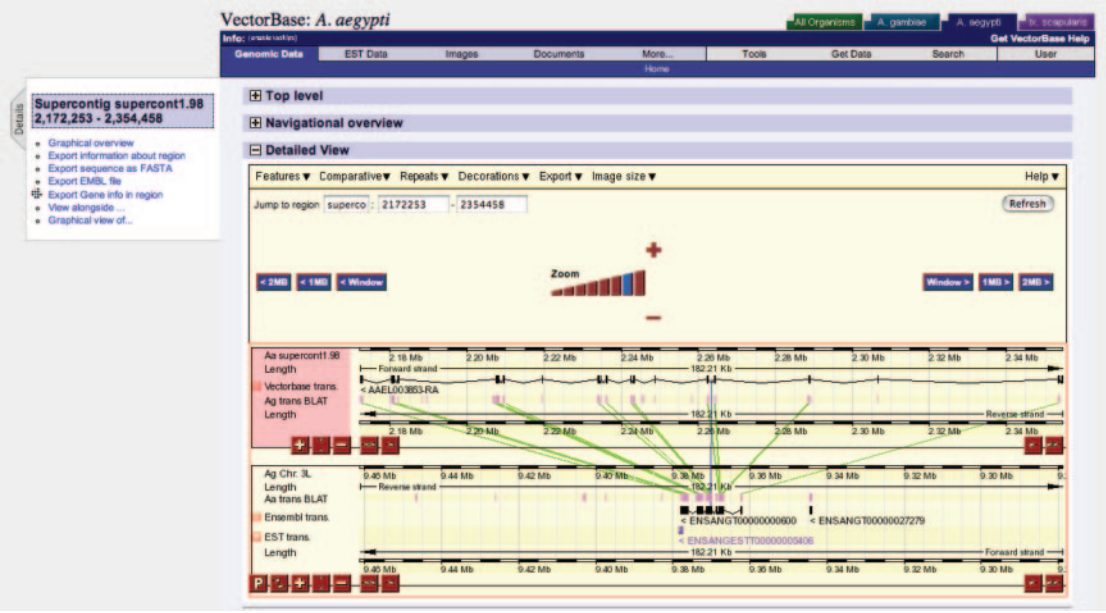

Figure 1. Comparative display of *Aedes aegypti* gene AAEL003853 (upper panel) with the *Anopheles gambiae* orthologue (lower panel). Green lines join blocks of similarity between the two genomes and highlights the expansion of intron size due to an increased frequency of repeat sequences.

FUTURE DIRECTIONS

At least four new arthropod vector genomes will soon be incorporated into VectorBase: the mosquito *Culex pipiens quinquefasciatus*, the tick *Ixodes scapularis*, the kissing bug *Rhodnius prolixus* and the human body louse, *Pediculus humanus* (see Table 1 for more details). Furthermore, the genomes of two molecular forms of *A.gambiae* (the S and M forms which are considered to be incipient or possibly distinct species) will soon be completed and integrated into VectorBase. We will continue to re-annotate the existing mosquito genomes to improve gene prediction drawing more on manual/community annotation and comparative analysis with additional arthropod genomes.

The increased importance of manual/community annotation is being addressed by the development of a CHADO-based database for tracking internal VectorBase manual annotation and submissions from the community.

Other material of interest to the vector community is being incorporated, including the newly developed controlled vocabulary of mosquito anatomy (http://obo.sourceforge.net/detail.cgi?mosquito_anatomy) and other vector-related ontologies.

VectorBase is an ongoing project and the scope and usability of the site are improving rapidly. The coming year will see a significant expansion in the number of vector genomes housed.

ACKNOWLEDGEMENTS

The core VectorBase project is funded by contract HHSN266200400039C from the NIAID, and supported, in part, by the BioMalPar network of excellence. We would like to thank the reviewers for their helpful comments and insights. Funding to pay the Open Access publication charges for this article was provided by NIH-NIAID.

Conflict of interest statement. None declared.

REFERENCES

1. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498–511.
2. Berriman,M., Ghedin,E., Hertz-Fowler,C., Blandin,G., Renauld,H., Bartholomeu,D.C., Lennard,N.J., Caler,E., Hamlin,N.E., Haas,B. *et al.* (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science*, **309**, 416–422.
3. Ivens,A.C., Peacock,C.S., Worthey,E.A., Murphy,L., Aggarwal,G., Berriman,M., Sisk,E., Rajandream,M.A., Adlem,E., Aert,R. *et al.* (2005) The genome of the kinetoplastid parasite, Leishmania major. *Science*, **309**, 436–442.
4. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
5. Krasprzyk,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 acces to biological data. *Genome Res.*, **14**, 160–169.
6. Ribeiro,J.M.C., Topalis,P. and Louis,C. (2004) AnoXcel: an *Anopheles gambiae* protein database. *Insect Mol. Biol.*, **13**, 449–457.
7. Holt,R.A., Subramanian,G.M., Halpern,A., Sutton,G.G., Charlab,R., Nusskern,D.R., Wincker,P., Clark,A.G., Ribeiro,J.M., Wides,R. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.
8. Curwen,V., Eyras,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.
9. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
10. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
11. Kriventseva,E.V., Koutsos,A.C., Blass,C., Kafatos,F.C., Christophides,G.K. and Zdobnov,E.M. (2005) Toward *A.gambiae* functional genomics. *Genome Res.*, **15**, 893–899.
12. Li,J., Riehle,M.M., Zhang,Y., Xu,J., Oduol,F., Gomez,S.M., Eigelmeier,K., Ueberheide,B.M., Shabanowitz,J., Hunt,D.F. *et al.* (2006) *Anopheles gambiae* genome reannotation through synthesis of *ab initio* and comparative gene prediction algorithms. *Genome Biol.*, **7**, R24.
13. Kalume,D.E., Peri,S., Reddy,R., Zhong,J., Okulae,M., Kumar,N. and Pandey,A. (2005) Genome annotation of *Anopheles gambiae* using mass spectrometry-derived data. *BMC Genomics*, **19**, 128.