General Information

The Eukaryotic Pathogen, Vector and Host Informatics Resource (VEuPathDB) is one of two Bioinformatics Resource Centers (BRCs) funded by the US National Institute of Allergy and Infectious Diseases (NIAID), with additional support from the Wellcome Trust (UK).

These resources stem from support initially provided for the Plasmodium Genome Database by the Burroughs Wellcome Fund (2000-2) and a research grant from NIAID (2002-6). The BRC program was initiated in 2004 to provide public access to computational platforms and analysis tools enabling collection, management, integration and mining of genomic information and other large-scale datasets relevant to infectious disease pathogens including their interaction with mammalian hosts and invertebrate vectors of disease. Two BRCs are currently funded:

- VEuPathDB focuses on eukaryotic pathogens and invertebrate vectors of infectious diseases, , encompassing data from prior BRCs devoted to parasitic species (EuPathDB), fungi (FungiDB) and vector species (VectorBase).
- PATRIC & VIPR focus on bacterial and viral pathogens.

VEuPathDB has also received funding from the Wellcome Trust (UK), Bill & Melinda Gates Foundation, and US Department of Agriculture to support informatics efforts focusing on kinetoplastida and fungal organisms with special emphasis on improving functional annotation for select genome sequences and families of genes.

VEuPathDB provides access to diverse genomic and other large scale datasets related to eukaryotic pathogens and invertebrate vectors of disease (see Data Summary). Organisms supported by this resource include (but are not limited to) the NIAID list of emerging and re-emerging infectious diseases.

Component web sites are constructed using a common infrastructure and standard data analysis and loading procedures, allowing the use of VEuPathDB as a single point of entry for each (or all) of these community resources, and the opportunity to leverage orthology for searches across taxa.

- Genome Info & Stats provides a list of all organisms available in this website.
- Data Sets provides a list of all information in this website integrated into VEuPathDB, with relevant references.

Current Funding

The Eukaryotic Pathogen, Vector and Host Informatics Resources (VEuPathDB) is funded by the National Institute of Allergy and Infectious Diseases (NIH/DHHS) under Contract No. NIH HHS 75N93019C00077.

VEuPathDB also receives funding from the Wellcome Trust (UK) to support informatics efforts focusing on kinetoplastida and fungal organisms with special emphasis on improving functional annotation of genomes. Grant numbers: 212929/Z/18/Z and 218288/Z/19/Z.

Data Access Policy

All data on these websites are provided freely for public use, through the contributions of many researchers involved in generating genome sequences, functional genomics datasets, and additional information. These data often derive from ongoing research, and are not guaranteed to be accurate. When using data obtained from VEuPathDB, it is important to cite the original publications and contributors. Please see Citing VEuPathDB.

Citing VEuPathDB in Publications and Presentations

If you use a VEuPathDB resource, we invite you to please cite the most relevant publication. This PubMed filter provides a list of the most recent VEuPathDB publications.

Please note that much of the data in VEuPathDB is provided by independent researchers, please cite them if you use their data. See Data Sets for a list of all information integrated into VEuPathDB, and related publications.

For acknowledgements in presentations, you may wish to use one or more of the following logos (right/control click to copy):

Additional resources leveraging the same infrastructure: MicrobiomeDB and ClinEpiDB.

Website Privacy Policy

We do not use or share any of your personal information for any purpose unrelated to the functionality of the websites; however, we do collect some information to help us understand how our sites are being used and to improve community support. See Website Privacy Policy.

About OrthoMCL

Current Release 6.3

In this release, 1 **Peripheral** species was added. Thus, OrthoMCL now predicts orthology for a total of 564 organisms (414 **Peripheral** and 150 **Core** organisms). Proteins from the 1 new species were mapped into **Core** ortholog groups. Then, a new set of **Residual** ortholog groups (e.g. OG6r3_101799) were formed from the collection of all unmapped **Peripheral** proteins. See below for the methods.

To see the current set of organisms as well as their proteome sources and orthology statistics, go to Proteome Sources and Statistics.

Downloads: Go to the download site to obtain the protein sequences and ortholog groups used in this release.

Method for Forming and Expanding Ortholog Groups

Proteins are placed into Ortholog Groups by the following steps:

1. The OrthoMCL algorithm (see below) is employed on proteins from a set of 150 **Core** species to form **Core** ortholog groups. These species were carefully chosen based on proteome quality and widespread placement across the tree of life. Each Core protein is placed by the algorithm into a **Core** ortholog group consisting of one or more proteins. Core group names have the format OG6_xxxxxx (e.g., OG6_101327). OG6 refers to OrthoMCL release 6; for each sub-release (e.g., 6.1, 6.2, etc), the Core species and the Core ortholog group names will remain constant.
2. The proteins from hundreds of additional organisms, termed **Peripheral** organisms, are mapped into the Core groups. To do this, NCBI BLASTP is used to compare each Peripheral protein to each Core protein in the Core groups. (Note that Peripheral proteins that were previously added to the Core group are NOT used in the BLASTP.) Then, each **Peripheral** protein is assigned to the Core group containing the Core protein with the best BLAST score, but only if the E-Value is <1e-5 and the percent match length is >=50%.
3. All Peripheral proteins that fail to map to a Core group are collected and subjected to independent OrthoMCL analysis, forming **Residual** groups consisting of one or more proteins. Residual group names have the format OG6r1_xxxxxx (e.g., OG6r1_101327), where OG6 refers to release 6 and r1 refers to sub-release 1.
4. For each subsequent sub-release (which will occur every ~3 months along with other VEuPathDB sites), proteomes from additional **Peripheral** organisms will be processed as in steps 2 and 3 above. However, step 3 will differ slightly because the previous set of Residual groups will be disassembled, leaving the previous unmapped Peripheral proteins to be combined with the new unmapped Peripheral proteins. All of these proteins will be used to form new Residual groups (e.g., OG6r2_xxxxxx).
5. On occasion, the set of Core species will be re-defined, as more appropriate proteomes become available. In this case, new Core groups (e.g., OG7_xxxxxx) and Residual groups (e.g., OG7r1_xxxxxx) will be formed.

This design allows for the addition of proteomes at every sub-release (e.g., 6.1, 6.2, etc). Note that **Core** groups (e.g., OG6_101327) will remain between sub-releases, though these groups will expand as Peripheral proteins are mapped in. In contrast, **Residual** groups will exist only for that sub-release; thus, Residual groups are useful in allowing the user to find proteins related to their protein(s) of interest, but are not stable groups.

The OrthoMCL Algorithm

See the **OrthoMCL Algorithm Document** for a detailed description of the OrthoMCL algorithm.

In overview:

- All-v-all BLASTP of the proteins.
- Compute *percent match length*
  - Select whichever is shorter, the query or subject sequence. Call that sequence S.
  - Count all amino acids in S that participate in any HSP.
  - Divide that count by the length of S and multiply by 100
- Apply thresholds to blast result. Keep matches with E-Value < 1e-5 percent match length >= 50%
- Find potential inparalog, ortholog and co-ortholog *pairs* using the Orthomcl Pairs program. (These are the pairs that are counted to form the *Average % Connectivity* statistic per group.)
- Use the MCL program to cluster the pairs into groups

Background on Orthology and Prediction

Orthologs are homologs separated by speciation events. Paralogs are homologs separated by duplication events. Detection of orthologs is

becoming much more important with the rapid progress in genome sequencing (Glover et al. 2019).

OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. It provides not only groups shared by two or more species/genomes, but also groups representing species-specific gene expansion families. Thus, it serves as an important utility for automated eukaryotic genome annotation. OrthoMCL starts with reciprocal best hits within each genome as potential in-paralog/recent paralog pairs and reciprocal best hits across any two genomes as potential ortholog pairs. Related proteins are interlinked in a similarity graph. Then, MCL (Markov Clustering algorithm; Dongen 2000; www.micans.org/mcl) is invoked to split mega-clusters. This process is analogous to the manual review in COG construction. MCL clustering is based on weights between each pair of proteins, so to correct for differences in evolutionary distance the weights are normalized before running MCL.

OrthoMCL is similar to the INPARANOID algorithm (Remm et al. 2001) but is extended to cluster orthologs from multiple species. OrthoMCL clusters are coherent with groups identified by EGO (Lee et al. 2002), and an analysis using EC number suggests a high degree of reliability (Li et al. 2003).

We evaluated the performance of seven widely-used orthology detection algorithms that use three general prediction strategies: phylogeny-based, evolutionary distance-based and BLAST-based (Chen, et al. 2007). Specifically, we used Latent Class Analysis (LCA), a statistical technique appropriate for testing large data sets when no gold standard is available. Our results show an overall trade-off between sensitivity and specificity among these algorithms, with INPARANOID and OrthoMCL performing best with False Positive (FP) and False Negative (FN) error rates lower than 20%.

Software

OrthoMCL was originally implemented by Li Li. The software was not available for download.

Version 1.4 was developed as publicly available software by Feng Chen (This version is now not supported).

Version 2.0 was re-engineered to handle large-scale datasets (hundreds of genomes) by Steve Fischer, Mark Heiges, John Iodice, and Ryan Thibodeau.

Publications

1. Li Li, Christian J. Stoeckert, Jr., and David S. Roos
   OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes
   Genome Res. 2003 13: 2178-2189. [Abstract] [Full Text]
2. Feng Chen, Aaron J. Mackey, Christian J. Stoeckert, Jr., and David S. Roos
   OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups
   Nucleic Acids Res. 2006 34: D363-8. [Full Text]
   * Please cite this paper if you publish research results benefited from OrthoMCL-DB.
3. Feng Chen, Aaron J. Mackey, Jeroen K. Vermunt, and David S. Roos
   Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes
   PLoS ONE 2007 2(4): e383. [Full Text]
   * Recommended in Faculty1000
4. Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., Shanmugam, D., Roos, D. S. and Stoeckert, C. J.
   Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups
   Current Protocols in Bioinformatics. 2011 35:6.12.1â€“6.12.19. [Full Text]

Acknowledgements

This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400037C. The major PIs are Drs. David Roos and Chris Stoeckert.

The OrthoMCL-DB project was initiated by Feng Chen in April 2005, and people from VEuPathDB, PCBI, and the Penn Center for Bioinformatics who have contributed to the project include: Mark Hickman, Steve Fischer, Brian Brunk, Omar Harb, Ryan Doherty, Aaron Mackey, Praveen Chakravarthula, Jerric Gao, and Charles Treatman. We'd also like to thank students and postdocs from the Roos lab for valuable suggestions, specifically Lucia Peixoto, and Dhanasekaran Shanmugam.

VEuPathDB Infrastructure Document

This document provides a high-level overview of the software infrastructure utilized by the VEuPathDB BRC to load, integrate and provide data to users. Please check a list of all the data sets loaded in our VEuPathDB sites utilizing this infrastructure.

Browser Compatibility Statement

We recognize that our users access VEuPathDB using various Internet Browsers and Operating Systems. Our goal is to ensure that you have the best possible experience on VEuPathDB, but it is impossible to develop applications that work identically, efficiently and effectively on all web browsers.

Based on our site usage statistics we support the following browsers used by greater than 95% of our visitors:

- Firefox
- Safari
- Chrome

Feel free to contact us about any browsing issues you might come across.

Data Loading and Database Schema

We use the **G**enomics **U**nified **S**chema (GUS) database schema and data loading infrastructure and its framework available at GusAppFramework. This includes not only a comprehensive database schema for integrating and representing genomic and functional (or post) genomic data but also tools for loading said data into that system. We have made some extensions to the schema and tools for VEuPathDB specific purposes primarily to generate de-normalized views of the data for query optimization purposes.

Our data are all stored in Oracle12c databases. Our software infrastructure also supports PostgreSQL but we have some Oracle specific SQL constructs in our model that would need to be changed in order to run successfully in PostgreSQL.

We load all data using an in house engineered workflow system called ReFlow. Briefly, ReFlow is engineered to be an efficient graph-based workflow system. In it each step (node in the graph) has the ability to be undone and subsequently rerun with updated data. This was a significant requirement as it enables us to undo entire genomes when the annotation or underlying sequence changes. This results in automated removal of all data dependent on that genome. When the step is re-run with the new annotation, all dependent data are recomputed and reloaded automatically, thus greatly improving our ability to keep these complex databases up-to-date.

The ReFlow workflow system utilizes another piece of software developed at the University of Pennsylvania to schedule, manage and monitor running tasks called DistribJob. DistribJob distributes tasks generated from a large input dataset such as a set of sequences to compute nodes in a cluster for analysis and retrieves and collates the results in an efficient manner. We automate the running of large compute tasks on compute clusters located at the University of Pennsylvania and the University of Georgia.

▶ **>>> Github Repositories**

Web Presentation System and User Interfaces

Our websites are based on code that we developed and have released to the community called the Strategies-WDK (Strategies Web Development Kit) which enables the graphical strategies search system. You can download the software and see documentation for this toolkit at Strategies-WDK. This toolkit enables us to represent our data as an XML model which is then turned into the web interfaces that are presented to users using these tools. As we develop new tools and services, we are transitioning towards a cloud-ready Restful architecture with user workspaces accessing tools in containerized services.

▶ **>>> Github Repositories**

System Hardware and Third Party Software

VEuPathDB maintains redundant database and content web servers at the University of Pennsylvania and the University of Georgia to minimize interruptions for our users during maintenance periods. Additionally VEuPathDB compute and data loading servers are located at the University of Pennsylvania.

Server configurations are coordinated and deployed through Puppet automation software (http://puppetlabs.com/). Custom infrastructure software is versioned and deployed through standard RPM/YUM mechanisms. When appropriate, software builds are automated using Jenkins Continuous Integration Server (http://jenkins-ci.org/)

System infrastructure statistics (CPU load, I/O, etc) are gathered with collectd (http://collectd.org/) and in-house applications and feed to Graphite (http://graphite.wikidot.com/) for human review. Nagios (http://www.nagios.org/) provides notifications of system degradations.

Both Universities also maintain large compute clusters that are heavily utilized by VEuPathDB in order to analyze and load incoming data in a timely fashion. The linked document below describes our actual hardware and includes a list of third-party software required in order to analyze, load and present data via our websites.

Community Interactions and Data Submission Policies

VEuPathDB serves a global scientific community that demands direct active support and community involvement. VEuPathDB outreach activities include:

- Organizing and running hands on training workshops and webinars (Google Map).
- Developing educational material in the form of exercises and online tutorials.
- Responding to support emails for users who contact us directly by clicking the "Contact Us" links in the header or footer of any VEuPathDB webpage (average response time is 48 hours).
- Holding open community meetings/forums with our diverse user base. These meetings are held in person at scientific conferences or using an online conferencing platform.
- Attending national and international meetings with active participation in the form of posters, presentations or help desks.
- Authoring peer reviewed manuscripts.
- Maintaining active social media presence in the form of a FaceBook page and Twitter feed.
- Providing a clear data handling and release policy to investigators to encourage submission of data prepublication.

How to submit data to VEuPathDB

The Eukaryotic Pathogen, Vector & Host Informatics Resources (https://VEuPathDB.org) is a Bioinformatics Resource Center (BRC) operated under contract from the US National Institute of Allergy and Infectious Diseases (NIAID) and the Wellcome Trust. VEuPathDB is charged with ensuring that genomic (and other large-scale) datasets pertaining to supported pathogens are conveniently accessible to the worldwide community of biomedical researchers. This document summarizes policies associated with releasing datasets on VEuPathDB. Our goal is to

help the communities that we serve ensure that their data are FAIR, Findable, Accessible, Interoperable and Reusable.

VEuPathDB welcomes submissions of genomic-scale data concerning eukaryotic microbes, fungi, vectors of human disease, and host-pathogen interactions. The VEuPathDB contract from NIAID provides support for biosecurity pathogens, including *Babesia*, *Cryptosporidium*, *Entamoeba*, *Giardia*, *Microsporidi*a (various genera), *Toxoplasma*, *Plasmodium*, and related taxa (*Acanthamoeba*, *Gregarina*, *Neospora*, *Theileria*) and also arthropod vectors (ticks, mites, mosquitoes, kissing bugs, tsetse flies, sand flies, lice, etc.) of human disease, as well as a snail that serves as an intermediate host, and comparator species. Support for kinetoplastid parasites (*Crithidia*, *Endotrypanum*, *Leishmania*, *Trypanosoma*) is provided by The Wellcome Trust. The FungiDB project encompasses a large (and growing) number of species supported by both NIAID and the Wellcome Trust. Please contact us if you have data from other species that should be incorporated into VEuPathDB!
Please review our Data Submission Policy.

Our most common data types include transcriptomics, proteomics, metabolomics, epigenomics, population-level and isolate information. In one form or another, VEuPathDB currently represents datasets in the following categories:

- Sequence (genomic [nuclear and organellar])
- ESTs and RNA-seq, generated on various platforms)
- Host-response data
- Comparative genomic information
- DNA polymorphism and population genetics data
- Sequences and metadata pertaining to field and clinical isolates and collections (with geo-spatiotemporal and other metadata)
- Chromatin modification data (ChIP-chip and ChIP-seq)
- Manually curated and automatically generated gene models and other annotation (GO terms, InterPro domains, etc.)
- Transcript and proteomic profiling
- Host response data sets (multiple platforms)
- Interactome data
- Protein structural information
- Metabolic pathways and metabolomics data
- Phenotype information, reagents (clones, antibodies, etc.)
- Publication references
- Image data, etc.

We also accept other genomic-scale data and are open to suggestions. Use the Contact Us link to make suggestions. We look forward to working with you!

How to submit data for integration in VEuPathDB

To submit your data for integration, fill out the appropriate VEuPathDB Dataset nomination form listed below. If your data cannot be submitted via our forms, use the Contact Us link to send a brief description (two or three sentences) of your data.

Genomes & high throughput sequencing data (e.g. RNA-Seq, ChiP-Seq, isolates typed by Whole Genome Sequencing or by sequencing limited genetic loci) must be available in The International Nucleotide Sequence Database Collaboration (INSDC) such as NCBI GenBank, EMBL-EBI ENA or DDBJ.

Once the dataset is prioritised for loading, we will export the data directly from INSDC. Note: while genome sequences must be available in INSDC, functional annotation (e.g. gene names, GO terms, etc.) can be submitted directly to VEuPathDB.

Tell us about your data as early as possible, to allow ample time for scheduling into VEuPathDB release cycles. Depending on the dataset type, we can provide instructions on how to transfer your data to us (e.g. formats of proteomics datasets may differ depending on the nature and scale of the data to be transferred), or we may be able to facilitate data submission to a repository (e.g. GenBank, GEO/ArrayExpress, etc.).

VectorBase resource: Gene manual annotations (change of exon-intron boundaries, creation of new genes) and metadata (gene names/symbols and functional description) can be submitted via Apollo (Coming soon). If you submit a gene annotation before you submit a manuscript for publication, we can generate GeneIDs that can be linked out to within the publication. Gene deletions are not handled via Apollo, please Contact Us with supporting evidence. Metadata can also be submitted by sending a spreadsheet file for batch submissions, follow this link for information about the 12 columns heading that are required. To add publications to a gene send us the corresponding PubMed link.

- ▶ **Genome Sequence and/or Annotation**
- ▶ **High Throughput or Next Generation Sequencing** – RNA, DNA or ChIP Sequencing
- ▶ **Microarray**
- ▶ **Proteomics**
- ▶ **Quantitative Proteomics**
- ▶ **ChIP-chip**
- ▶ **Isolates typed by sequencing limited genetic loci**
- ▶ **Isolates or Strains typed by High Throughput Sequencing**
- ▶ **Bulk user comments form**
- ▶ **Phenotype collection form**

General Data Submission – Use the Contact Us form to tell us about data that does not fit any of the above categories

Related sites

- Previous EuPathDB Workshops
- Companion
- OrthoMCL
- GeneDB
- ModBase at UCSF
- Tetrahymena Genome
- The Arabidopsis Information Resource
- NAR Database Summary Paper Categories

VEuPathDB Publications and Citations

View **publications that cite us** in Google Scholar

This PubMed filter provides a current list of the most recent publications about VEuPathDB resources

---

(Book Chapter)
Omar Harb, Jessica C. Kissinger and David S. Roos on behalf of the EuPathDB group
ToxoDB: the functional genomic resource for *Toxoplasma*
Toxoplasma gondii 3rd edition
Edited by Louis Weiss and Kami Kim (2020)
https://doi.org/10.1016/B978-0-12-815041-2.00023-2

(Book Chapter)
Susanne Warrenfeltz and Jessica Kissinger on behalf of the EuPathDB Consortium
Accessing *Cryptosporidium* omic & isolate data via CryptoDB.org
Methods in Molecular Biology
Editor, Jan Mead (2020)
https://doi.org/10.1007/978-1-4939-9748-0_22

(Book Chapter)
Omar S. Harb and David S. Roos on behalf of the EuPathDB Consortium
ToxoDB: Functional Genomics Resource for *Toxoplasma* and Related Organisms
Methods in Molecular Biology
Editor, Christopher J. Tonkin (2020)
https://doi.org/10.1007/978-1-4939-9857-9_2

(Book Chapter)
Susanne Warrenfeltz, Evelina Y Basenko, Kathryn Crouch, Omar S. Harb, Jessica C. Kissinger, Achchuthan Shanmugasundram and Fatima Silva-Franco
EuPathDB: the eukaryotic pathogen genomics database resource
Methods in Molecular Biology
Editor, Martin Kollmar (2018)
https://doi.org/10.1007/978-1-4939-7737-6_5

Evelina Y. Basenko, Jane A. Pulman, Achchuthan Shanmugasundram, Omar S. Harb, Kathryn Crouch, David Starns, Susanne Warrenfeltz, Cristina Aurrecoechea, Christian J. Stoeckert, Jr., Jessica C. Kissinger, David S. Roos and Christiane Hertz-Fowler
FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. (2018)
J. Fungi 2018, 4(1), 39
https://doi.org/10.3390/jof4010039

Francislon S Oliveira, John Brestelli, Shon Cade, Jie Zheng, John Iodice, Steve Fischer, Cristina Aurrecoechea, Jessica C Kissinger, Brian P Brunk, Christian J Stoeckert, Jr, Gabriel R Fernandes, David S Roos, Daniel P Beiting
MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments
Nucleic Acids Research. 2017 doi.org/10.1093/nar/gkx1027

Cristina Aurrecoechea; Ana Barreto; Evelina Y. Basenko; John Brestelli; Brian P. Brunk; Shon Cade; Kathryn Crouch; Ryan Doherty; Dave Falke; Steve Fischer; Bindu Gajria; Omar S. Harb; Mark Heiges; Christiane Hertz-Fowler; Sufen Hu; John Iodice; Jessica C. Kissinger; Cris Lawrence; Wei Li; Deborah F. Pinney; Jane A. Pulman; David S. Roos; Achchuthan Shanmugasundram; Fatima Silva-Franco; Sascha Steinbiss; Christian J. Stoeckert Jr; Drew Spruill; Haiming Wang; Susanne Warrenfeltz; Jie Zheng
EuPathDB: the eukaryotic pathogen genomics database resource
Nucleic Acids Research 2017 doi: 10.1093/nar/gkw1105

Mitraka, E., Topalis, P., Dritsou, V., Dialynas, E., & Louis, C.
Describing the breakbone fever: IDODEN, an ontology for dengue fever
PLoS neglected tropical diseases, 3 February 2015; 9(2), e0003479
https://doi.org/10.1371/journal.pntd.0003479

Warren, A. S., Aurrecoechea, C., Brunk, B., Desai, P., Emrich, S., Giraldo-CalderÃ³n, G. I., Harb, O., Hix, D., Lawson, D., Machi, D., Mao, C., McClelland, M., Nordberg, E., Shukla, M., Vosshall, L. B., Wattam, A. R., Will, R., Yoo, H. S., & Sobral, B.
RNA-Rocket: an RNA-Seq analysis resource for infectious disease research
Bioinformatics, 1 May 2015; 31(9), 1496â€“1498
https://doi.org/10.1093/bioinformatics/btv002

Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, Ho N, Gesing S; VectorBase Consortium, Madey G, Collins FH, Lawson D.
VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases
Nucleic Acids Res. 2015 Jan;43(Database issue):D707-13
https://doi.org/10.1093/nar/gku1117

Harb OS, Roos DS.
The Eukaryotic Pathogen Databases: a functional genomic resource integrating data from human and veterinary parasites
Methods Mol Biol. 2015;1201:1-18
[Full text]

Dugan VG, Emrich SJ, Giraldo-CalderÃ³n GI, Harb OS, Newman RM, Pickett BE, Schriml LM, Stockwell TB, Stoeckert CJ Jr, Sullivan DE, Singh I, Ward DV, Yao A, Zheng J, Barrett T, Birren B, Brinkac L, Bruno VM, Caler E, Chapman S, Collins FH, Cuomo CA, Di Francesco V, Durkin S, Eppinger M, Feldgarden M, Fraser C, Fricke WF, Giovanni M, Henn MR, Hine E, Hotopp JD, Karsch-Mizrachi I, Kissinger JC, Lee EM, Mathur P, Mongodin EF, Murphy CI, Myers G, Neafsey DE, Nelson KE, Nierman WC, Puzak J, Rasko D, Roos DS, Sadzewicz L, Silva JC, Sobral B, Squires RB, Stevens RL, Tallon L, Tettelin H, Wentworth D, White O, Will R, Wortman J, Zhang Y, Scheuermann RH.
Standardized metadata for human pathogen/vector genomic sequences
PLoS One. 2014 Jun 17;9(6):e99979
[Full text | PDF]

Topalis, P., Mitraka, E., Dritsou, V., Dialynas, E., & Louis, C.
IDOMAL: the malaria ontology revisited
Journal of biomedical semantics, 13 September 2013; 4(1), 16
https://doi.org/10.1186/2041-1480-4-16

Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Cade S, Doherty R, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Hu S, Iodice J, Kissinger JC, Kraemer ET, Li W, Pinney DF, Pitts B, Roos DS, Srinivasamoorthy G, Stoeckert CJ Jr, Wang H, Warrenfeltz S.
EuPathDB: the eukaryotic pathogen database
Nucleic Acids Res. 2013 Jan;41(Database issue):D684-91
[Full text | PDF]

Zerlotini A, Aguiar ER, Yu F, Xu H, Li Y, Young ND, Gasser RB, Protasio AV, Berriman M, Roos DS, Kissinger JC, Oliveira G.
SchistoDB: an updated genome resource for the three key schistosomes of humans
Nucleic Acids Res. 2013 Jan;41(Database issue):D728-31
[Full text | PDF]

Megy, K., Emrich, S. J., Lawson, D., Campbell, D., Dialynas, E., Hughes, D. S., Koscielny, G., Louis, C., Maccallum, R. M., Redmond, S. N., Sheehan, A., Topalis, P., Wilson, D., & VectorBase Consortium
VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics
Nucleic acids research, 1 January 2012; 40(Database issue), D729â€“D734
https://doi.org/10.1093/nar/gkr1089

Stajich JE, Harris T, Brunk BP, Brestelli J, Fischer S, Harb OS, Kissinger JC, Li W, Nayak V, Pinney DF, Stoeckert CJ Jr, Roos DS.
FungiDB: an integrated functional genomics database for fungi
Nucleic Acids Res. 2012 Jan;40(Database issue):D675-81
[Abstract | Full text | PDF]

Topalis, P., Dialynas, E., Mitraka, E., Deligianni, E., Siden-Kiamos, I., & Louis, C.
A set of ontologies to drive tools for the control of vector-borne diseases
Journal of biomedical informatics, February 2011; 44(1), 42â€“47
https://doi.org/10.1016/j.jbi.2010.03.012

Maccallum, R. M., Redmond, S. N., & Christophides, G. K.
An expression map for *Anopheles gambiae*
BMC genomics, 20 December 2011; 12, 620
https://doi.org/10.1186/1471-2164-12-620

Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ Jr.
Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups
Curr. Protoc. Bioinform. 35:6.12.1-6.12.19. Â© 2011 by John Wiley & Sons, Inc. PMID: 21901743
[Abstract | Full text | PDF]

Aurrecoechea C., Barreto A, Brestelli J., Brunk BP., Caler EV., Fischer S., Gajria B., Gao X., Gingle A, Grant G., Harb OH., Heiges M., Iodice J., Kissinger JC., Kraemer ET., Li W., Nayak V., Pennington C., Pinney DF., Pitts B., Roos DS., Srinivasamoorthy G., Stoeckert CJ Jr., Treatman C., and Wang H.
AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species
Nucleic Acids Research. 39(Database issue):D612-9, 2011 Jan
https://doi.org/10.1093/nar/gkq1006

Fischer S, Aurrecoechea C, Brunk BP, Gao X, Harb OS, Kraemer ET, Pennington C, Treatman C, Kissinger JC, Roos DS, Stoeckert CJ.
The Strategies WDK: a graphical search interface and web development kit for functional genomics databases
Database (Oxford). 2011 Jun 23;2011:bar027. Print 2011. PMID: 21705364 PMCID: PMC3122067
[Abstract | Full text | PDF]

Topalis, P., Mitraka, E., Bujila, I., Deligianni, E., Dialynas, E., Siden-Kiamos, I., Troye-Blomberg, M., & Louis, C.
IDOMAL: an ontology for malaria
Malaria journal, 10 August 2010; 9, 230
https://doi.org/10.1186/1475-2875-9-230

H. Wang, J. DeBarry, J. C. Kissinger
SynView - Synteny Browser with GBrowse Integration
Editor, Gabriel P.C. Fung (2010)
A Practical Guide to Bioinformatics Analysis- 2010 iConcept Press Ltd

---

Aurrecoechea C. Brestelli J. Brunk BP. Fischer S. Gajria B. Gao X. Gingle A. Grant G. Harb OS. Heiges M. Innamorato F. Iodice J. Kissinger JC. Kraemer ET. Li W. Miller JA. Nayak V. Pennington C. Pinney DF. Roos DS. Ross C. Srinivasamoorthy G. Stoeckert CJ Jr. Thibodeau R. Treatman C. Wang H.
EuPathDB: a portal to eukaryotic pathogen databases
Nucleic Acids Research 2010 38(Database issue):D415-9
[Abstract | Full text | PDF]

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, Gardner MJ, Gingle A, Grant G, Harb OS, Heiges M, Hertz-Fowler C, Houston R, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Logan FJ, Miller JA, Mitra S, Myler PJ, Nayak V, Pennington C, Phan I, Pinney DF, Ramasamy G, Rogers MB, Roos DS, Ross C, Sivam D, Smith DF, Srinivasamoorthy G, Stoeckert CJ Jr., Subramanian S, Thibodeau R, Tivey A, Treatman C, Velarde G, Wang H.
TriTrypDB: a functional genomic resource for the Trypanosomatidae
Nucleic Acids Research 2010 38(Database issue):D457-D462; doi:10.1093/nar/gkp851
[Abstract | Full text | PDF]

Dialynas, E., Topalis, P., Vontas, J., & Louis, C.
MIRO and IRbase: IT tools for the epidemiological monitoring of insecticide resistance in mosquito disease vectors
PLoS neglected tropical diseases, 23 June 2009; 3(6), e465
https://doi.org/10.1371/journal.pntd.0000465

Megy, K., Hammond, M., Lawson, D., Bruggner, R. V., Birney, E., & Collins, F. H.
Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase
Infection, genetics and evolution, May 2009; 9(3), 308â€“313
https://doi.org/10.1016/j.meegid.2007.12.007

Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., Campbell, K. S., Christophides, G. K., Christley, S., Dialynas, E., Hammond, M., Hill, C. A., Konopinski, N., Lobo, N. F., MacCallum, R. M., Madey, G., Megy, K., Meyer, J., Redmond, S., Severson, D. W., â€¦ Collins, F. H.
VectorBase: a data resource for invertebrate vector genomics
Nucleic acids research, 1 January 2009; 37(Database issue), D583â€“D587
https://doi.org/10.1093/nar/gkn857

Aurrecoechea, C., J. Brestelli, B. P. Brunk, J. Dommer, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, J. A. Miller, V. Nayak, C. Pennington, D. F. Pinney, D. S. Roos, C. Ross, C. J. Stoeckert, Jr., C. Treatman, and H. Wang
PlasmoDB: a functional genomic database for malaria parasites
Nucleic Acids Res. 2009. 37:D539-D543
https://doi.org/10.1093/nar/gkn814

Aurrecoechea, C., J. Brestelli, B. P. Brunk, J. M. Carlton, J. Dommer, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, J. A. Miller, H. G. Morrison, V. Nayak, C. Pennington, D. F. Pinney, D. S. Roos, C. Ross, C. J. Stoeckert, Jr., S. Sullivan, C. Treatman, and H. Wang
GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*
Nucleic Acids Res. 2009. 37:D526-D530
https://doi.org/10.1093/nar/gkn631

Topalis, P., Tzavlaki, C., Vestaki, K., Dialynas, E., Sonenshine, D.E., Butler, R., ... & Louis, C.
Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase
Insect molecular biology, 30 January 2008; 17(1), 87-9
https://doi.org/10.1111/j.1365-2583.2008.00781.x

Topalis, P., Lawson, D., Collins, F.H., & Louis, C.
How can ontologies help vector biology
Trends in parasitology, 1 June 2008; 24(6), 249-52
https://doi.org/10.1016/j.pt.2008.03.002

Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., Campbell, K. S., Christophides, G. K., Christley, S., Dialynas, E., Emmert, D., Hammond, M., Hill, C. A., Kennedy, R. C., Lobo, N. F., MacCallum, M. R., Madey, G., Megy, K., Redmond, S., Russo, S., â€¦ Collins, F. H.
VectorBase: a home for invertebrate vectors of human pathogens
Nucleic acids research, 1 January 2007; 35(Database issue), D503â€“D505
https://doi.org/10.1093/nar/gkl960

B. Gajria, A. Bahl, J. Brestelli, J. Dommer, S. Fischer, X. Gao, M. Heiges, J. Iodice, J. C. Kissinger, A. J. Mackey, et al.
ToxoDB: an integrated *Toxoplasma gondii* database resource
Nucleic Acids Res. 2007. 36:D553-6
https://doi.org/10.1093/nar/gkm981

Cristina Aurrecoechea, Mark Heiges, Haiming Wang, Zhiming Wang, Steve Fischer, Philippa Rhodes, John Miller, Eileen Kraemer, Christian J. Stoeckert, Jr., David S. Roos and Jessica C. Kissinger
ApiDB: integrated resources for the apicomplexan bioinformatics resource center
Nucleic Acids Research. 2007. 35:D427-30
https://doi.org/10.1093/nar/gkl880

Christian J. Stoeckert Jr., Steve Fisher, Jessica Kissinger, Mark Heiges, Cristina Aurrecoechea, Bindu Gajria, David S. Roos
PlasmoDB v5: New Looks, New Genomes
Trends in Parasitology. 2006. 22(12): 543-546
[Abstract]

Patricia L. Whetzel, Shailesh V. Date, Kobby Essien, Martin J. Fraunholz, Bindu Gajria, Gregory R. Grant, John Iodice, Jessica C. Kissinger, Philip T. Labo, Arthur J. Milgram, David S. Roos, and Christian J. Stoeckert Jr.
PlasmoDB: The Plasmodium Genomics and Functional Genomics Resource
In silico Genomics and Proteomics: Functional Annotation of Genomes and Proteins. Nicola Mulder and Rolf Apweiler (eds.). Nova Science Publishers. 2006

Haiming Wang, Yanqi Su, Aaron J. Mackey, Eileen T. Kraemer and Jessica C. Kissinger
SynView: A GBrowse-compatible Approach to Visualizing Comparative Genome Data
Bioinformatics. 2006. 22(18):2308-2309
[Abstract | Full text |PDF | Supplement]

Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, Rhodes P, Wang S, He CZ, Su Y, Miller J, Kraemer E, Kissinger JC.
CryptoDB: a *Cryptosporidium* bioinformatics resource update
Nucleic Acids Res. 2006 Jan 1;34:D419-22
[Abstract | Full text | PDF]

Feng Chen, Aaron J. Mackey, Christian J. Stoeckert Jr. and David S. Roos
OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups
Nucleic Acids Res. 2006 Jan 1;34:D363-8
[Abstract | Full text | PDF]

Kriventseva, E. V., Koutsos, A. C., Blass, C., Kafatos, F. C., Christophides, G. K., & Zdobnov, E. M.
AnoEST: toward *A. gambiae* functional genomics
Genome research, 17 May 2005; 15(6), 893â€“899
https://doi.org/10.1101/gr.3756405

Stoeckert CJ Jr.
Functional genomics databases on the web
Cell Microbiol. 2005 7(8), 1053-9
https://doi.org/10.1111/j.1462-5822.2005.00553.x

J.C. Kissinger and D.S. Roos
Getting the most out of bioinformatics resources
In: Malaria Parasites, A.P. Waters & C.J. Janse, editors. Horizon, Norfolk UK. 2004
https://www.caister.com/hsp/abstracts/mal/03.html

Ribeiro, J.M., Topalis, P., & Louis, C.
AnoXcel: an Anopheles gambiae protein database
Insect molecular biology, 16 September 2004; 13(5), 449-57
https://doi.org/10.1111/j.0962-1075.2004.00503.x

VEuPathDB Data Analysis Methods

VEuPathDB draws data from many sources. To facilitate comparisons across data sets, we analyze all data with standardized, data type-specific analyses. All data of one type are analyzed with the same workflow. Although our results may show some differences from an authorâ€™s publication, our re-analysis of the data makes it feasible to compare data sets from very different sources and to update the data analysis with contemporary methods. For transparency, the methods we use to analyze data are presented here.

Contents

Genome analyses

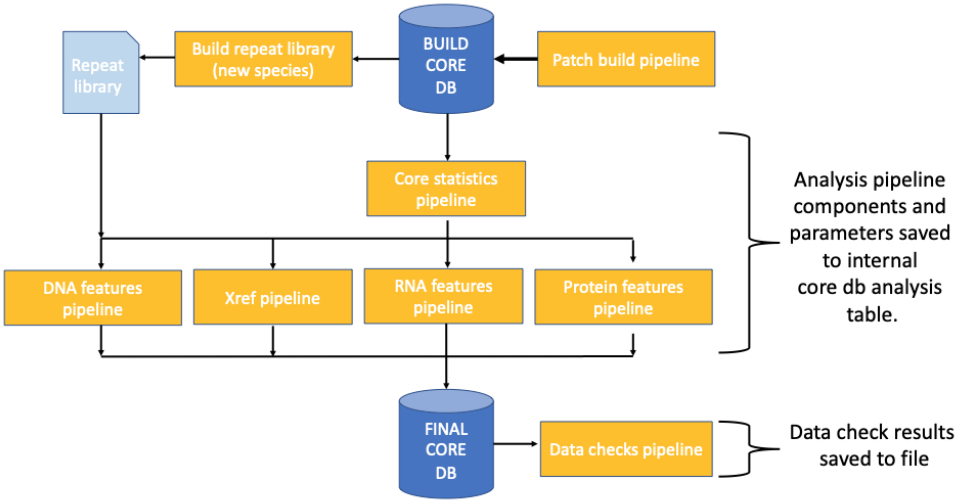
Genome sequence and annotation are analyzed by the EBI Pipeline supplemented with three in-house analyses. In the rare case that the EBI pipeline cannot be applied to a genome, we use a series of in-house analyses in lieu of the EBI Pipelines.

EBI Pipeline

VEuPathDB employs the Ensembl genome analysis for analyzing genomic sequence to enhance annotations. While most of the genomic sequence (FASTA) are integrated into VEuPathDB from an INSDC repository, genome annotation (GFF3) may come from either the INSDC repository or a community submission.

Core database pipelines (figure 1)- Primary genomic sequence and structural annotation data are loaded into a core database and run through 6 pipelines: core statistics, DNA feature annotation, external cross reference annotation, RNA gene annotation, repeat feature annotation, and protein feature annotation. The main pipelines applied to the core database and their components are listed in table 1.

Configuration details for each pipeline are determined in Ensembl hive pipeline config files for each pipeline. Since the Ensemble pipelines may change to accommodate bioinformatic advances, pipeline component programs (e.g. Interpro for protein features), versions, and parameters are recorded in the core db analysis table. Final data check results are saved to disk and manually reviewed to determine if the final core db is suitable for release to be loaded into the GUS system.



Core database analysis pipelines and hive components

Example ehive pipelines, modules, programs and parameter data from coredb analysis table

Supplements to the EBI Pipelines

VEuPath DB supplements the EBI pipeline with workflows that produce data for EST alignments, Open reading frames, and synteny (Table).

EST alignments: BLAT is applied to EST sequences that have been blocked using RepeatMasker.

Open reading frame generation: Open reading frames are generated from genomic DNA or EST sequences. The analysis produces a gff file containing the ORFs (50 or more amino acid translations of the input nucleic acid) for the 6 reading frames. The translations in all 6 reading frames do not necessarily begin with MET but always end at a stop codon. ORF names are in the form template-frame-start-end, e.g. AAEL01000396-5-5847-4366.

Synteny: VEuPathDB uses an in-house script called runMercator to run pair-wise alignments that employs Mercator and MAVID for comparative genome analysis. Mercator generates orthology maps using genomes and exon coordinates to create exon translations for protein BLAT alignments. The orthology maps are used as a guide for MAVID which also uses a phylogenetic newick tree to generate gene alignments.

Details for the supplements to the EBI pipelines

In-house genome analyses in Lieu of the EBI Pipeline

On rare occasions the EBI pipeline cannot be applied to a genome. For example, genomes that are not housed at an INSDC repository cannot be analyzed by the EBI pipeline. VEuPathDB uses the following in-house analyses in lieu of the EBI pipeline.

BLAT against NRDB: For every genome, VEuPathDB runs BLAT alignements of the annotated proteins against the GenBank Non-Redundant Protein Sequence Database (NRDB) to identify possible relationships and alignments outside the scope of VEuPathDB-supported organisms.

Compute open reading frames: Open reading frames are generated from genomic DNA or EST sequences. The analysis produces a gff file containing the ORFs (50 or more amino acid translations of the input nucleic acid) for the 6 reading frames. The translations in all 6 reading frames do not necessarily begin with MET but always end at a stop codon. ORF names are in the form template-frame-start-end, e.g. AAEL01000396-5-5847-4366.

DNA repeats: The Tandem Repeats Finder program locates and displays tandem repeats in genomic sequences.

EST alignments: BLAT is applied to EST sequences that have been blocked using RepeatMasker.

Protein domain annotations: InterProScan scans protein sequences against the protein signatures of the InterPro member databases and generates a file containing the domain matched, description of the InterPro entry, GO descriptions and E-values.

Signal peptide prediction: Signal P is used to identify signal peptides and their likely cleavage sites. A signal peptide is a short peptide present at the N-terminus of most newly synthesized proteins that are destined towards the secretory pathway.

Syntenic sequences: VEuPathDB uses an in-house script called runMercator to run pair-wise alignments that employs Mercator and MAVID for comparative genome analysis. Mercator generates orthology maps using genomes and exon coordinates to create exon translations for protein BLAT alignments. The orthology maps are used as a guide for MAVID which also uses a phylogenetic newick tree to generate gene alignments.

Transmembrane domain prediction: TMHMM is used to predict transmembrane domain presence and topology from protein sequences.

tRNA gene prediction: tRNAScan identifies transfer RNA genes in transcript or genome sequences.

Details for the VEuPathDB in-house pipelines

Orthology

Orthologs are homologs separated by speciation events. Paralogs are homologs separated by duplication events. Identification of orthologs is important in annotating newly-sequenced genomes, in predicting gene function and in understanding gene evolution. Orthology relationships are predicted across all VEuPathDB sites using OrthoMCL, a VEuPathDB website designed to predict, analyze, and display orthology relationships.

OrthoMCL

OrthoMCL employs BLAST and clustering to predict orthologs and paralogs amongst all organisms at VEuPathDB sites, as well as additional organisms spread across Eukaryota, Bacteria, and Archaea. In the OrthoMCL algorithm, each protein is assigned an ortholog group (such as OG6_135465 or OG6r2_106455), which contains proteins predicted to be orthologs or paralogs of each other. The OrthoMCL website allows users to explore ortholog groups and the proteins within these groups. The site offers the ability to explore protein sequences, sequence alignments, PFam domains, EC numbers, and sequence similarity metrics. More detailed information is provided on the About OrthoMCL page. The following sections describe how OrthoMCL predictions are used throughout VEuPathDB sites.

Orthology on the gene page

A gene page, like that for *Plasmodium falciparum* PF3D7_1371700, contains an 'Orthology and synteny' section. This section displays the 'Ortholog Group' (with a link to the group page on OrthoMCL), 'Orthologs and Paralogs within PlasmoDB' (listing genes within the same ortholog group), 'Strains summary' (listing the corresponding gene in strains of the same species), and 'Synteny' (a JBrowse view of syntenic chromosomes with shading connecting genes in the same OrthoMCL-predicted group).

Function prediction on the gene page

Orthology is used to predict gene function on the gene pages. We and others have found that orthologs and paralogs share a high degree of function. A gene page, like that for *Plasmodium falciparum* PF3D7_1371700, contains a 'Function prediction' section. The EC Numbers table may contain EC numbers that have been predicted for the gene, based on other genes within the same ortholog group having been assigned this EC number. This EC number prediction in turn influences the 'Pathways and interactions' section, containing the 'Metabolic Pathways' and 'Metabolic Pathways Reactions' tables. These tables list the predicted pathways and reactions for a gene assigned a specific EC Number. Thus, a user can make inferences about a gene's function in specific pathways based on orthology.

Searches for genes based on orthology

There are three Searches that employ orthology to identify a set of genes. First, in the Orthology Phylogenetic Profile Search, users can identify genes that are found in certain taxonomic groups but not in others (i.e., that have a specific pattern of conservation across species). For example, users can identify genes that are present in *Toxoplasma gondii* but not present in mammals (i.e., the ortholog group is present in *T. gondii* but not in mammals). Second, in the Paralog Count Search, users can identify genes that have a specific range of paralogs within a species (i.e., genes that have undergone an expansion in a species due to one or more gene duplications). Third, within a Search Strategy (see tutorial here), users can choose to transform their list of genes from one species into the list of orthologs in another species. This is particularly useful when a user is working with a species that is little studied. For example, a user has found a set of genes induced by hypoxia in *Aspergillus fumigatus* and wants to know their orthologs in the less-studied *Aspergillus oryzae*.

Proteomics

VEuPathDB integrates the results of proteomics experiments as peptides aligned to a reference genome or as abundance data assigned to a gene. We do not reanalyze the raw mass spec data but instead use an in-house plugin that loads found peptides or abundance data from tab delimited input files of a specific format.

Details for the VEuPathDB in-house proteomics pipeline

RNA-Sequence

VEuPathDB integrates RNA-Seq data from many different experiments and analyzes all data with the same EBI RNA-Seq analysis pipeline. The RNA sequence data that we integrate is processed at EBI.

The following is a general outline of the analysis process.

- Trim poor quality data (Trimmomatic)
- HiSAT2 alignment to a reference genome
- HT-Seq-count to tally aligned reads per gene
- Convert to transcripts per kilobase million (TPM)
- DESeq2 to determine differential expression

EBI RNA-Seq pipeline details

ChIP-Sequence

VEuPathDB integrates ChIP-Seq data from many different experiments and sources. DNA seq data are aligned to the reference genome using Bowtie2. Alignment results are converted to bigwig and displayed in JBrowse.

Copy Number Variation

VEuPathDB uses coverage from whole genome sequencing data to estimate gene and chromosome copy numbers in sequenced strains. The bowtie2 alignments generated during SNP analysis are used as a starting point. HTseq-count is used to count the number of reads that align to each gene and the values are converted to transcripts per million (TPM). Assuming that the median TPM value represents a single copy gene on a chromosome of constitutive ploidy, we can infer gene or chromosome duplications by comparing the TPM values for individual genes or the median TPM for individual chromosomes to the whole genome median using custom scripts based on the method described in PMID: 22038252. Additionally, coverage is calculated in 1 kb bins across the genome, normalised to the constitutive ploidy and converted to bigwig format for visualisation in JBrowse.

Haploid number and gene dose are metrics used to define copy number in VEuPathDB. Haploid number is the number of genes on an individual chromosome. Gene dose is the total number of genes in an organism, accounting for copy number of the chromosome. For example, a single-copy gene in a diploid organism has a haploid number of 1 and a gene dose of 2.

Searches for genes based on Copy Number Variation

There are two searches that query copy number data in VEuPathDB. The first, Identify Genes based on Copy Number (CNV) returns genes that are present at copy numbers within a range that you specify. The search can be configured to return genes based on the haploid number or gene dose. The second search, Identify Genes based on Copy Number Comparison (CNV), returns genes for which the copy number varies between the reference and your chosen isolates. This search compares the estimated copy number of a gene in the resequenced strain(s) with the copy number in the reference annotation. The copy number in the reference annotation is calculated as the number of genes that are both on the same chromosome and in the same ortholog group as the gene of interest. We infer that these genes have arisen as a result of tandem duplication of a common ancestor. In this search, the metric for copy number is the haploid number, which is the number of copies of a gene on a single chromosome.

Genetic Variation and SNP calling

VEuPathDB analyzes whole genome resequencing data to call single nucleotide polymorphisms of isolates. The method employed by VEuPathDB to call SNPs from short read sequencing like Illumina reads, follows these steps:

- Reads are aligned to the reference genome using bowtie2
- The resulting BAM file from bowtie2 is sorted and a pileup file using samtools is generated
- Reads around indels are realigned using GATK
- SNPs and indels are called consensus sequence using VarScan is generated:
  - P value <= 0.01
  - minimum aligned reads >= 5
  - minimum read frequency >= 0.8
- SNP calls where coverage is >2.5x the median coverage are removed to limit erroneous calls in repeat regions
- At each SNP position â€œlike referenceâ€ calls are generated for each strain that is identical to the reference to give the full picture of each SNP

Microarray data

VEuPathDB integrates microarray data from high density oligonucleotide as well as spotted arrays. In general, the data comes to us as intensities associated with probes. VEuPathDB does not reanalyze the original fluorescence data. We process the data we receive according to the following outline:

- Map the array probes to the reference genome's transcriptome
- Filter the data to remove outliers.
- Normalize
  - For one channel data we perform a robust multi-array average (RMA) normalizations.
  - For two channel data we perform a Loess normalization
- Compute the average probe intensity per gene.
- Compute the expression average per gene.
  - First, average the technical replicates.
  - Second, average the biological replicates (if any).
- Optional: perform differential expression analysis if there is a sufficient number of biological replicates.

Protein Array data

VEuPathDB integrates protein array data from serm antibody microarray experiments. In general, the data comes to us as intensities associated with probes. VEuPathDB does not reanalyze the original fluorescence data. Although each experiment and data set can have special considerations, we process the data according to the following outline:

- Map the array probes to the reference genome's transcriptome
- Filter the data to remove outliers.
- Normalize
  - For one channel data we perform a robust multi-array average (RMA) normalizations.
  - For two channel data we perform a Loess normalization
- Compute the average probe intensity per gene.
- Compute the expression average per gene.
  - First, average the technical replicates.
  - Second, average the biological replicates (if any).
- Optional: perform differential expression analysis if there is a sufficient number of biological replicates.

Metabolic Pathways

VEuPathDB integrates metabolic pathways from KEGG and MetaCyc. For TriTrypDB, pathways are also integrated from LeishCyc and TrypanoCyc. Metabolic pathways are associated with genes via Enzyme Commission annotations.

VEuPathDB Infrastructure Document

This document provides a high-level overview of the software infrastructure utilized by the VEuPathDB BRC to load, integrate and provide data to users. Please check a list of all the data sets loaded in our VEuPathDB sites utilizing this infrastructure.

Browser Compatibility Statement

We recognize that our users access VEuPathDB using various Internet Browsers and Operating Systems. Our goal is to ensure that you have the best possible experience on VEuPathDB, but it is impossible to develop applications that work identically, efficiently and effectively on all web browsers.

Based on our site usage statistics we support the following browsers used by greater than 95% of our visitors:

- Firefox
- Safari
- Chrome

Feel free to contact us about any browsing issues you might come across.

Data Loading and Database Schema

We use the **G**enomics **U**nified **S**chema (GUS) database schema and data loading infrastructure and its framework available at GusAppFramework. This includes not only a comprehensive database schema for integrating and representing genomic and functional (or post) genomic data but also tools for loading said data into that system. We have made some extensions to the schema and tools for VEuPathDB specific purposes primarily to generate de-normalized views of the data for query optimization purposes.

Our data are all stored in Oracle12c databases. Our software infrastructure also supports PostgreSQL but we have some Oracle specific SQL constructs in our model that would need to be changed in order to run successfully in PostgreSQL.

We load all data using an in house engineered workflow system called ReFlow. Briefly, ReFlow is engineered to be an efficient graph-based

workflow system. In it each step (node in the graph) has the ability to be undone and subsequently rerun with updated data. This was a significant requirement as it enables us to undo entire genomes when the annotation or underlying sequence changes. This results in automated removal of all data dependent on that genome. When the step is re-run with the new annotation, all dependent data are recomputed and reloaded automatically, thus greatly improving our ability to keep these complex databases up-to-date.

The ReFlow workflow system utilizes another piece of software developed at the University of Pennsylvania to schedule, manage and monitor running tasks called DistribJob. DistribJob distributes tasks generated from a large input dataset such as a set of sequences to compute nodes in a cluster for analysis and retrieves and collates the results in an efficient manner. We automate the running of large compute tasks on compute clusters located at the University of Pennsylvania and the University of Georgia.

▶ **>>> Github Repositories**

Web Presentation System and User Interfaces

Our websites are based on code that we developed and have released to the community called the Strategies-WDK (Strategies Web Development Kit) which enables the graphical strategies search system. You can download the software and see documentation for this toolkit at Strategies-WDK. This toolkit enables us to represent our data as an XML model which is then turned into the web interfaces that are presented to users using these tools. As we develop new tools and services, we are transitioning towards a cloud-ready Restful architecture with user workspaces accessing tools in containerized services.

▶ **>>> Github Repositories**

System Hardware and Third Party Software

VEuPathDB maintains redundant database and content web servers at the University of Pennsylvania and the University of Georgia to minimize interruptions for our users during maintenance periods. Additionally VEuPathDB compute and data loading servers are located at the University of Pennsylvania.

Server configurations are coordinated and deployed through Puppet automation software (http://puppetlabs.com/). Custom infrastructure software is versioned and deployed through standard RPM/YUM mechanisms. When appropriate, software builds are automated using Jenkins Continuous Integration Server (http://jenkins-ci.org/)

System infrastructure statistics (CPU load, I/O, etc) are gathered with collectd (http://collectd.org/) and in-house applications and feed to Graphite (http://graphite.wikidot.com/) for human review. Nagios (http://www.nagios.org/) provides notifications of system degradations.

Both Universities also maintain large compute clusters that are heavily utilized by VEuPathDB in order to analyze and load incoming data in a timely fashion. The linked document below describes our actual hardware and includes a list of third-party software required in order to analyze, load and present data via our websites.

Community Interactions and Data Submission Policies

VEuPathDB serves a global scientific community that demands direct active support and community involvement. VEuPathDB outreach activities include:

- Organizing and running hands on training workshops and webinars (Google Map).
- Developing educational material in the form of exercises and online tutorials.
- Responding to support emails for users who contact us directly by clicking the "Contact Us" links in the header or footer of any VEuPathDB webpage (average response time is 48 hours).
- Holding open community meetings/forums with our diverse user base. These meetings are held in person at scientific conferences or using an online conferencing platform.
- Attending national and international meetings with active participation in the form of posters, presentations or help desks.
- Authoring peer reviewed manuscripts.
- Maintaining active social media presence in the form of a FaceBook page and Twitter feed.
- Providing a clear data handling and release policy to investigators to encourage submission of data prepublication.

VEuPathDB Website Privacy Policy

UPDATED: April 12, 2020

Introduction

VEuPathDB (also referred to as â€˜weâ€™ throughout this document) is committed to protecting its usersâ€™ data and privacy. The purpose of this page is to provide you with information about how the data we collect from users of VEuPathDB websites is used or shared. We may update this Privacy Notice from time to time. We encourage you to visit this page frequently and take note of the date updated field above.

We do not use or share any of your personal information for any purpose unrelated to the functionality of the websites; however, we do collect some information to help us understand how our sites are being used in order to improve community support and to enhance the VEuPathDB communityâ€™s experience when visiting our sites.

Information Automatically Collected

When you browse VEuPathDB sites, certain information about your visit will be collected. We automatically collect and store the following type of information about your visit:

- The IP address of the client making the request. Often the IP address is that of your personal computer or smart phone; however, it might be that of a firewall or proxy your internet provider manages.
- The operating system and information about the browser used when visiting the site.
- The date and time of each visit.
- Pages visited.
- The address of a referring page. If you click a link on a website that directs you to a VEuPathDB page, the address of that originating web page will be collected. This â€œreferrerâ€ information is transmitted as part of the browser and server communications; it is not based on any marketing or partnering agreements with the referring site.

This automatically collected information does not identify you personally unless you include personally identifying information in a support form request; see the â€œContact Usâ€ policy below for details. We use this information to measure the number of visitors to our site. The aggregate data may be included in prospectuses and reports to funding agencies.

Information You Directly Provide

The Basket, Favorites, Public Strategies, Gene/Sequence Comment and GBrowse Track features of the VEuPathDB websites require that you register for an account. A valid email address is required so we can send you your temporary account password. An anonymous email service can be used if you do not want to provide personally identifying information.

Your email address will be used to send you infrequent alerts if you subscribe to receive them. We do not sell or distribute email addresses to third parties.

We also ask for your name and institution during account registration. If you add a comment to a Gene or a Sequence, your name and institution will be displayed with the comment. If you make one of your strategies public, your name and institution will be displayed with it. We do not routinely verify the validity of names and institutions associated with comments or public strategies; however, we will delete accounts or comments if we believe them to be fraudulent based on inappropriate activity or posted content. We will not sell or distribute your name or institution to third parties.

When you log in, the client IP address is recorded. This IP address can be correlated with the address automatically collected as noted above. If your user profile personally identifies you, then it may be possible to associate you with your detailed activity on VEuPathDB web sites.

â€œContact Usâ€ Form

The header on each web page includes a â€œContact Usâ€ link to a form where users can submit questions, error reports, feature requests, and dataset proposals. Submissions through this form are emailed to VEuPathDB staff and recorded in a project management application accessible only by VEuPathDB staff.

The form includes a field for an email address. If the email address identifies you personally, say if you use your institutional email, then your correspondence with us will likewise be linked to you. A valid email is not strictly required, although we cannot reply to you without one.

When you submit the form, your IP address and browser version will be recorded for internal use. In the case of reported bugs or other site errors, this information may be used by technical staff to help locate your session in the server logs to aid in troubleshooting the issue. This does have the side effect of making it possible to associate an IP address with an email address which may, in turn, personally identify you. However, VEuPathDB does not publicly release this information.

How VEuPathDB Uses Cookies

VEuPathDB uses cookies to associate multiple requests by your web browser into a stateful session. Cookies are essential to track the state of query strategies, gene baskets and authentication.

Some cookies persist only for a single session. The information is recorded temporarily and is erased when the user quits the session or closes the browser. Others may be persistently stored on the hard drive of your computer until you manually delete them from a browser folder or until they expire, which can be months after they were last used.

Cookies can be disabled in your browser (refer to your browserâ€™s documentation for instructions); however, the majority of the website functionality will be unavailable if cookies are disabled.

Google Analytics

Google Analytics provides aggregate measurements of website traffic including counts of page hits and unique users along with statistics on countries of origin.

The raw measurements and statistics are only available to approved VEuPathDB staff. Aggregated data may be included in prospectuses and reports to funding agencies.

Third-Party Websites and Applications

Third-party websites and applications are not exclusively operated or controlled by VEuPathDB. By using these third-party websites, individuals may be providing nongovernmental third-parties with access to personally identifying information.

Twitter

VEuPathDB maintains a presence on Twitter in the form of a VEuPathDB branded page. This page allows for a direct connection with end users to promote information related VEuPathDB services and to disseminate educational information on research publications, news and events related to the biology of eukaryotic pathogens. Postings may also include information about planned service maintenance and outages.

Twitter collects profile information such as name and email address about users who register to use this third-party website. Depending on the userâ€™s privacy settings, this information may be displayed on the userâ€™s profile page or in the userâ€™s tweets which may be retweeted on VEuPathDBâ€™s page. The VEuPathDB Twitter account may post the authors and institutions of publicly published scientific papers and news articles. VEuPathDB does not actively collect or maintain personally identifying information through its use of Twitter. VEuPathDB will redact or refrain from retweeting a posting that contains obviously identifiable personal information. A Twitter account is not required to read VEuPathDB postings on Twitter. VEuPathDB does not collect or use personal information outside of Twitterâ€™s site.

Twitter is hosted and maintained by a third party which may use browser tracking and related technologies to collect information about visitors to twitter.com and its affiliates. Refer to Twitterâ€™s privacy statement, https://twitter.com/en/privacy, for more information.

Facebook

VEuPathDB maintains a presence on Facebook in the form of a VEuPathDB branded page. This page allows for a direct connection with end users to promote information related VEuPathDB services and to disseminate educational information on research publications, news and events related to the biology of eukaryotic pathogens. Postings may also include information about planned service maintenance and outages.

Like Twitter, Facebook collects profile information, including name and email address, from its users. Depending on the userâ€™s privacy settings, this information may be displayed on the userâ€™s profile page along with any activity such as comments or â€œlikesâ€ on the VEuPathDB Facebook page or in posts that VEuPathDB shares on Facebook. VEuPathDB does not collect or use any personally identifying information outside of our Facebook page. To understand how Facebook collects and uses personal information, refer to their data policy page, https://www.facebook.com/policy.php.

YouTube

VEuPathDB maintains a presence on YouTube in the form of a VEuPathDB branded page. This page provides tutorials on the use of our websites.

YouTube also requires some information when users create an account, including an email address, and users may choose to provide a name and other identifying information in their public profile. Depending on their individual privacy settings, some personally identifiable information may be available to other users, including VEuPathDB. However, VEuPathDB does not collect or use any of that information outside of its YouTube interactions. You can view videos without signing in to an account, but you must be a registered user in order to comment. As YouTube is a Google service, the VEuPathDB youtube channel is subject to Googleâ€™s privacy policy.

Globus Genomics

The VEuPathDB Galaxy Data Analysis Service is a workspace for large-scale data analyses. Developed in partnership with Globus Genomics, workspaces offer a private analysis platform with published workflows and pre-loaded annotated genomes. The workspace is accessed through the â€œAnalyze My Experimentâ€ tab on the home page of any VEuPathDB resource and can be used to upload your own data, compose and run custom workflows, retrieve results and share workflows and data analyses with colleagues.

The VEuPathDB Galaxy Data Analysis Service is hosted by Globus Genomics, an affiliate of Globus. The first time you visit VEuPathDB Galaxy you will be asked to sign up with Globus in order to set up your private Galaxy workspace. Linking your Globus account with your VEuPathDB account is necessary so that input data and analysis results can be transferred between the two systems. We encrypt data transfers and storage, but ultimately, we cannot guarantee the security of data transmissions among VEuPathDB, Globus and affiliates, Amazon Cloud Services, and the user. It is your responsibility to back up your data and obtain any required permissions from your study and/or institution prior to uploading data for analyses on the VEuPathDB Galaxy platform. Do not use, transmit, upload or share any human identifiable information in the files you analyze. VEuPathDB, Globus and affiliates, the University of Georgia, the University of Pennsylvania, the University of Liverpool, and Amazon Cloud Services do not take any responsibility and are not liable for the loss and/or release of any data you analyze via the VEuPathDB Galaxy platform. We encourage you to review the Globusâ€™ privacy policy.

Your Rights based on the General Data Protection Regulation (GDPR)

To read more about GDPR please check the GDPR website.

1. The right of transparency and modalities. The privacy policy should be clear and easy to follow in explaining what data we collect and how we use it.
2. The right to be informed about when data is gathered. This is described in the privacy policy, during the registration process (if you choose to register), site banner and an email sent out to all registered users on May 25, 2018.
3. The right of access. You can ask for what sepecific data we have about you and how we use it.
4. The right to rectification. We will correct any errors in your personal data that you point out to us.
5. The right to be forgotten. We are happy to delete your account and info when you make such a request.
6. The right to restrict processing. You have the right to request that we restrict the use of your data.
7. The right for notification obligation regarding rectification/erasure/restriction.
8. The right to data portability.
9. The right to object to the processing of your personal data at any time.
10. The right in relation to automated decision making and profiling. Basically, you have the right not to be subject to decisions based solely on automated processing which significantly affect you.

To make any of the above stated requests or if you have any questions please email us at help@VEuPathDB.org

VEuPathDB personnel

VEuPathDB Management

Beatrice Amos, Annotation Manager
Cristina Aurrecoechea, User Interface and Portal Manager
Bob Belnap, Systems and Databases Manager
John Brestelli, Data Development Manager
Brian Brunk, VEuPathDB Senior Manager
Andy Jones, Wellcome Trust PI; Co-I NIAID BRC Contract
George Christophides, Co-I, NIAID BRC Contract
Kathryn Crouch, Co-I, Wellcome Trust
Jeremy DeBarry, Project Coordinator
Steve Fischer, Software and Infrastructure Manager
Paul Flicek, Co-I, NIAID BRC Contract
Omar Harb, Director of Scientific Outreach & Education
Jessica C Kissinger, Joint-PI, NIAID BRC Contract; WT Co-PI
Dan Lawson, Project Coordinator
Wei Li, Data Loading Manager
Mary Ann McDowell, Joint-PI, NIAID BRC Contract
David S Roos, Joint-PI, NIAID BRC Contract; WT Co-PI
Chris J Stoeckert, Co-I, NIAID BRC Contract

To contact any one of us please use the contact us form.

Current VEuPathDB Team members

Beatrice Amos[4], Rachel Ankirskiy[1], Cristina Aurrecoechea[1], Matthieu Barba[9], Ana Barreto[3], Evelina Basenko[4], Wojtek Bazant[2], Dan Beiting[2], Bob Belnap[1], Ulrike Böhme[5], John Brestelli[3], Brian Brunk[2], Danielle Callan[2], Mikkel Christensen[9], George Christophides[8], Kathryn Crouch[6], Katie Cybulski[7], Jeremy DeBarry[1], Ryan Doherty[3], Yikun Duan[2], Dave Falke[1], Steve Fischer[3], Paul Flicek[9], Bindu Gajria[2], Gloria I. Giraldo-CalderÃ³n[7], Omar S. Harb[2], Elizabeth Harper[2], Danica Helb[2], Mark Hickman[2], Connor Howington[7], Sufen Hu[2], Jay Humphrey[1], John Iodice[3], Andy Jones[4], John Judkins[2], Sarah Kelly[8], Jessica C. Kissinger[1], Dae Kun Kwon[7], Kris Lamoureux[1], Daniel Lawson[8], Wei Li[2], Brianna Lindsay[2], Jamie Long[2], Bob MacCallum[8], Gareth Maslen[9], Mary Ann McDowell[7], Greg Milewski[2], Jarek Nabrzyski[7], David S. Roos[2], Samuel Rund[7], Steph Wever Schulman[2], Achchuthan Shanmugasundram[4], Vasili Sitnik[9], Drew Spruill[1], David Starns[4], Christian J. Stoeckert Jr.[3], Sheena Shah Tomko[2], Haiming Wang[1], Susanne Warrenfeltz[1], Robert Wieck[7], Paul Wilkinson[4], Mariann Winkelman[2], Lin Xu[2], Jie Zheng[3].

Previous VEuPathDB Team members, 2004-2020

Antelmo Aguilar[7], James Allen[9], Alexis Allot[9], Nora Besansky[7], Austin Billings[2], Sanjay Boddu[9], Steve Bogol[7], Ewan Birney[9], Andrew Brockman[8], Robert Bruggner[7], Ja'Shon Cade[3], Mark Caddick (Principal Investigator, WT 2019-2020)[4], David Campbell[7], Cristian Cocos[2], Frank Collins (VectorBase Principal Investigator, 2004-2018)[7], Kathy Couch[1], Greg Davies[7], Elaine Daugan[4], Ale Diaz Miranda[2], Emmanuel Dialynas, Jennifer Dommer[3], Vicky Dritsou, Scott Emrich[10], Xin Gao[2], William Gelbart[12], Sandra Gesing[7], Alan Gingle[1], Greg Grant[3], Matt Guidry[1], Martin Hammond[9], Mark Heiges[1], Christiane Hertz-Fowler (Principal Investigator, WT 2008-2019)[4], Nicholas Ho[8], Daniel Hughes[9], Frank Innamorato[3], San James[14], Amie Jaye[8], Fotis Kafatos[8], Paul Kersey[9], Ioannis Kimitzoglou[8], Nathan Konopinski[7], Carolyn Knoll[2], Eileen T. Kraemer[1], Nick Langridge[9], Cris Lawrence[2], Neil Lobo[7], Christos (Kitsos) Louis[11], Ross Madden[6], Greg Madey[7], Elisabetta Manduchi[3], Karine Megy[9], John A. Miller[6], Elvira Mitraka[11], Vishal Nayak[3], Cary Pennington[1], Deborah F. Pinney[3], Brian Pitts[1], Jane A.Pulman[4], Caleb Reinking[7], Seth Redmon, Chris Ross[1], Andrew Sheehan[7], Fatima Silva[4], Ganesh Srinivasamoorthy[1], Scott Szakonyi[7], Pantelis Toplais[11], Ryan Thibodeau[1], Charles Treatman[2], Betsy Wenthe[1], Matt Vander Werf[7], Maggie Werner-Washburne[13], Patricia L. Whetzel[3], Derek Wilson[9], Andrew Yates[9]

[1]University of Georgia, Athens, GA 30602, USA
[2]University of Pennsylvania, Philadelphia, PA 19104, USA
[3]University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA
[4]University of Liverpool, L69 3BX, UK
[5]Wellcome Sanger Institute, Hinxton, CB10 1RQ, UK
[6]Wellcome Centre for Integrative Parasitology, University of Glasgow, UK
[7]University of Notre Dame, Notre Dame, IN 46556, USA
[8]Imperial College London, South Kensington, London SW7 2BU, UK
[9]European Bioinformatics Institute, Hinxton, CB10 1SD, UK
[10]University of Tennessee, Knoxville, TN 37996, USA
[11]Institute of Molecular Biology and Biotechnology-FORTH, Heraklion, Crete, Greece

[12]Harvard University, Cambridge, MA 02138, USA
[13]University of New Mexico, Albuquerque, NM 87131, USA
[14]Makerere University and Infectious Diseases Research Collaboration (IDRC), Kampala, Uganda

VEuPathDB wishes to acknowledge

- All the community members who have contributed data (often pre-publication), entered user comments or sent us their suggestions.
- Scientists who provided images to be used as a template for the logos in our websites and for images used in the header section of our sites:

Â Â Â	Â Â Â **AmoebaDB:**Â Â Â	Â Â Â **MicrosporidiaDB:**Â Â Â	Â Â Â **ToxoDB:**Â Â Â	Â Â Â
Â Â Â	Â Â Â William PetriÂ Â Â	Â Â Â Gira BhabhaÂ Â Â	Â Â Â David RoosÂ Â Â	Â Â Â
Â Â Â	Â Â Â Serge AnkriÂ Â Â	Â Â Â Pattana Jaroenlak (Michael)Â Â Â	Â Â Â **TrichDB:**Â Â Â	Â Â Â
Â Â Â	Â Â Â Craig RobertsÂ Â Â	Â Â Â Damian EkiertÂ Â Â	Â Â Â Antonio Pereira-NevesÂ Â Â	Â Â Â
Â Â Â	Â Â Â Fiona HenriquezÂ Â Â	Â Â Â Michael CammerÂ Â Â	Â Â Â Marlene BenchimolÂ Â Â	Â Â Â
Â Â Â	Â Â Â Hugo Aguilar-DÃazÂ Â Â	Â Â Â **PiroplasmaDB:**Â Â Â	Â Â Â **TriTrypDB:**Â Â Â	Â Â Â
Â Â Â	Â Â Â Julio C. CarreroÂ Â Â	Â Â Â Ellen YehÂ Â Â	Â Â Â Rick TarletonÂ Â Â	Â Â Â
Â Â Â	Â Â Â **CryptoDB:**Â Â Â	Â Â Â Lowell KappemeyerÂ Â Â	Â Â Â Richard WheelerÂ Â Â	Â Â Â
Â Â Â	Â Â Â Boris StriepenÂ Â Â	Â Â Â Audrey LauÂ Â Â	Â Â Â Leandro Lemgruber SoaresÂ Â Â	Â Â Â
Â Â Â	Â Â Â **FungiDB:**Â Â Â	Â Â Â Dirk DobbelaereÂ Â Â	Â Â Â Margaret MullinÂ Â Â	Â Â Â
Â Â Â	Â Â Â ZyGoLifeÂ Â Â	Â Â Â Manoj DuraisinghÂ Â Â	Â Â Â Camila Silva GonÃ§alvesÂ Â Â	Â Â Â
Â Â Â	Â Â Â Jason StajichÂ Â Â	Â Â Â Brendan ElsworthÂ Â Â	Â Â Â Wanderley de SouzaÂ Â Â	Â Â Â
Â Â Â	Â Â Â Zachary LewisÂ Â Â	Â Â Â Caroline KeroackÂ Â Â	Â Â Â Maria Cristina Machado MottaÂ Â Â	Â Â Â
Â Â Â	Â Â Â **GiardiaDB:**Â Â Â	Â Â Â Isabelle CoppensÂ Â Â	Â Â Â	Â
Â Â Â	Â Â Â Fran GillinÂ Â Â	Â Â Â **PlasmoDB:**Â Â Â	Â Â Â	Â
Â Â Â	Â Â Â Tineke LauwaetÂ Â Â	Â Â Â Lawrence BannisterÂ Â Â	Â Â Â	Â
Â Â Â	Â Â Â Barbara DavidsÂ Â Â	Â Â Â Lewis TilneyÂ Â Â	Â Â Â	Â
Â Â Â	Â Â Â Scott DawsonÂ Â Â	Â Â Â Pedro MouraÂ Â Â	Â Â Â	Â

VEuPathDB Community Representatives

VEuPathDB encourages community members to provide feedback about our resources. We get feedback from many community members including those listed below who have been active in our open community meetings. We encourage you to get involved. Feel free to contact us any time.

Amoeba: Open community call and Carol Gilchrist, Upi Singh
Cryptosporidium: Gregory Buck, Guy Robinson, Karin Troell, Sumiti Vinayak, Jonathon Wastling, Giovanni Widmer, Lihua Xiao, Guan Zhu
Fungi: Bridget Barker, Elaine Bignell, Katherine Borkovich, Michael Bromley, Christina Cuomo, Tamara Doering, Jay Dunlap, Michael Freitag, Louise Glass, Kim Hammond-Kosack, Guilhem Janbon, Seogchan Kang, Theo Kirkland, Corby Kistler, Jennifer Lodge, Robin May, Jessie Uehling, Sinem Beyhan, Douglas Lake, Natalie Mitchell, Maureen Donlin, Vera Meyer, Marc Orbach, Nadia Ponts, Antonis Rokas, Jason Stajich, Matt Sachs, George R. Thompson, Martin Urban, Nathan P. Wiederhold
Giardia: Scott Dawson, Fran Gillin, Adrian Hehl, Aaron Jex, Hilary Morrison, John Samuelson, Cornelia Spycher, Staffan Svard
Microsporidia: James Becnel, Nicolas Corradi, Elizabeth Didier, Patrick Keeling, Emily Troemel, Louis Weiss
Piroplasma: Open community call and Choukri Mamoun
Plasmodium: John Adams, Chris Janse, Rays H.Y. Jiang, Shahid Khan, Stuart Ralph, Akhil Vaidya, Andy Waters
Toxoplasma: John Boothroyd, Jon P. Boyle, Vern B. Carruthers, Marc-Jan Gubbels, Kami Kim, Markus Meissner, Jeroen Saeij, Lilach Sheiner, Ross Waller, Michael White
Trichomonas: Jane Carlton, Patricia Johnson, Steven Sullivan, Jan Tachazy
Trypanosoma/Kinetoplastids: Fernan Aguero, Vivian Bellofatto, Richard Burchmore, George Cross, Angela Cruz, Antonio Estevez, Mark Field, Catarina Gadelha, Eva Gluenz,Keith Gull, John Kelly, Annette MacLeod, Jeremy Mottram, Torsten Ochsenreiter, Marc Ouellette, Barbara Papadopoulou, Laurie Read, Sergio Schenkman, Rick Tarleton, Brent Weatherly, Bill Wickstead, Michael (Mick) Urbaniak
Vectors: Gregory Dasch, Jeff Grabowski, MarÃa de Lourdes MuÃ±oz, Monika Gulia-Nuss, Sukanya Narasimhan, Kristin Michel, Michael Povelones, Igor Sharakhov, Ronald van Rij, Rob Waterhouse

Previous Scientific Working Group

VEuPathDB wishes to acknowledge previous scientific working group members. They provided regular feedback oversight and guidance.

Â	Â	Â	Â	Â	Â
Â Â Â	Â Â Â Lyric BartholomayÂ Â Â	Â Â Â Michael GottliebÂ Â Â	Â Â Â Malcolm McConvilleÂ Â Â	Â Â Â John TaylorÂ Â Â	Â Â Â
Â Â Â	Â Â Â Matt BerrimanÂ Â Â	Â Â Â Keith GullÂ Â Â	Â Â Â Nicola MulderÂ Â Â	Â Â Â Jake TuÂ Â Â	Â Â Â
Â Â Â	Â Â Â Bill BlackÂ Â Â	Â Â Â Matthew HahnÂ Â Â	Â Â Â Ull MunderlohÂ Â Â	Â Â Â Brett TylerÂ Â Â	Â Â Â
Â Â Â	Â Â Â John BoothroydÂ Â Â	Â Â Â Adrian HehlÂ Â Â	Â Â Â Daniel NeafseyÂ Â Â	Â Â Â Kenneth VernickÂ Â Â	Â Â Â

Â Â Â Â Â Â Greg BuckÂ Â Â	Â Â Â Steve HiggsÂ Â Â	Â Â Â Kenneth OlsonÂ Â Â	Â Â Â Sarah VolkmanÂ Â Â	Â Â Â
Â Â Â Â Â Â Geraldine ButlerÂ Â Â	Â Â Â Catherine HillÂ Â Â	Â Â Â Bill PetriÂ Â Â	Â Â Â Jonathan WastlingÂ Â Â	Â Â Â
Â Â Â Â Â Â Angela CruzÂ Â Â	Â Â Â Marcelo Jacobs-LorenaÂ Â Â	Â Â Â Barry PittendrighÂ Â Â	Â Â Â Scott WeaverÂ Â Â	Â Â Â
Â Â Â Â Â Â George DimopoulosÂ Â Â	Â Â Â Anthony (Tony) JamesÂ Â Â	Â Â Â Jeffrey PowellÂ Â Â	Â Â Â Louis WeissÂ Â Â	Â Â Â
Â Â Â Â Â Â Martin DonnellyÂ Â Â	Â Â Â Pedro Lagerblad de OliveiraÂ Â Â	Â Â Â Hillary RansonÂ Â Â	Â Â Â Dyann WirthÂ Â Â	Â Â Â
Â Â Â Â Â Â Patrick DuffyÂ Â Â	Â Â Â Greg LanzaroÂ Â Â	Â Â Â Alexander RaikhelÂ Â Â	Â Â Â Jennifer WortmanÂ Â Â	Â Â Â
Â Â Â Â Â Â Pascale GaudetÂ Â Â	Â Â Â Daniel MasigaÂ Â Â	Â Â Â Lincoln SteinÂ Â Â	Â Â Â Guiyun YanÂ Â Â	Â Â Â

VEuPathDB Glossary

Please check the NCBI glossary

3-Frame translation (forward)

Translation of a nucleotide sequence in all three possible reading frames in one direction, usually "on the top strand" of DNA.

3-Frame translation (reverse)

Translation of a nucleotide sequence in all three possible reading frames in the reverse direction, usually "on the bottom strand" of DNA.

AA sequence

Amino acid sequence.

Affymetrix genotyped SNP probes

Probes on Affymetrix SNP (single nucleotide polymorphism) arrays, which are used for SNP genotyping. See Affymetrix microarray technology and www.affymetrix.com

Affymetrix microarray technology

Microarray manufacturing technology developed by Affymetrix. Combines semiconductor fabrication techniques, solid phase chemistry, combinatorial chemistry, molecular biology, and robotics to generate a photolithographic manufacturing process in which oligonucleotides are synthesized directly on a chip. See www.affymetrix.com

Affymetrix probes

Probe on an Affymetrix microarray designed to determine whether or not the complementary sequence of RNA or DNA is present in a sample. Generally 25 nucleotides in length (25-mers), their short length provides higher specificity than longer probes. See Affymetrix microarray technology and www.affymetrix.com

Amitochondriate

Eukaryotic organism that lacks a mitochondrion. Examples include Giardia and other parasites such as Trachipleistophora and Entamoeba. However, most of these organisms contain what appear to be mitochondrial remnants as well as mitochondrial genes in their nuclear genomes.

Annotation

Identified feature within a sequence, such as a known or predicted gene, domain, motif, post-translational modification, etc.

Annotation density

Level to which a nucleotide or protein sequence has been annotated. See Annotation.

ApiCyc

Database/utility on EuPathDB used for searching and visualizing metabolic pathway information for organisms in EuPathDB; derivative database generated by analyzing various genomes (for example from Plasmodium, Cryptosporidium, and Toxoplasma) with SRI International's pathway tools.

ApiDots alignments

Consensus sequences found in the ApiDots database and generated by clustering and assembling Apicomplexan mRNA and EST sequences. These consensus sequences were subjected to database searches against protein and protein domain sequences.

Apicoplast

Nonphotosynthetic plastid found in almost all protozoan parasites belonging to the phylum Apicomplexa that have been examined. The apicoplast is surrounded by four membranes, giving rise to the theory that its presence in the Apicomplexa is the result of a secondary endosymbiosis (acquired by the engulfment of an ancestral alga and retention of the algal plastid). Similar to other endosymbiotic organelles (mitochondria, chloroplasts), the apicoplast contains its own genome as well as proteins that are encoded in the nucleus and post-translationally imported. The apicoplast is a vital organelle to the parasite's long-term survival.

Attribute

Inherent characteristic or feature; in a database, a data item related to a database object. For example, attributes of genes can include features such as introns and untranslated regions (UTRs).

BLAST

Basic local alignment search tool, a sequence similarity search tool used to quickly find local alignments between a query sequence and sequences in nucleotide or protein databases. Different versions of this search tool are available to match the types of query sequence and database used. See blastn, blastp, blastx, tblastn, and tblastx.

Boolean

System of logical thought developed by George Boole (1815-1864). In Boolean searching, an "and" operator between two words or values (for example, "apple AND orange") generates a search for items in a database containing both of the words or values. Similarly, an "or" operator between two words or values (for example, "apple OR orange") generates a search for items containing either word.

CDS

Coding sequence. Region of nucleotides that corresponds to the sequence of amino acids in a predicted protein and that includes start and stop codons. Unexpressed sequences (for example, the 5'-UTR, the 3'-UTR, and introns) are not included within a CDS. The CDS usually does not correspond to the actual mRNA sequence.

Centromere

Region of the chromosome or chromosomal structure essential for division and retention of the chromosome within the cell; point of a chromosome where the spindle fibers attach to pull the chromosome apart during cell division.

Chromosome

Macromolecule of DNA constituting the physical organization of DNA in a cell.

Coil

Three-dimensional spiral structure in protein macromolecules.

Contig

Contiguous genomic sequence assembled from overlapping primary sequences representing overlapping regions of a particular chromosome.

CryptoCyc

Database/utility built by analyzing the Cryptosporidium genome with SRI International's pathway tools; used for searching and visualizing Cryptosporidium metabolic pathway information.

Curated annotation

Annotation made under the supervision of a curator as opposed to a purely computational prediction. Curated predictions often

contain combinations of different types of evidence to support the annotation.

DNA/GC content

Content of guanine (G) and cytosine (C) in a fragment of DNA or a genome. Because GC pairs are more thermostable compared to the AT pairs, it was commonly believed that GC content played a vital part in adaptation to high temperatures, a hypothesis that has been refuted. In the genome browser, the DNA/GC content track displays a GC content graph of the reference sequence at low magnifications and the DNA sequence itself at higher magnifications.

Dalton

Unit of mass abbreviated Da and used to express atomic and molecular masses.

Deprecated gene

Genes with little or no evidence (similarities / expression) that overlapped (or were subsumed by) larger genes for which there was evidence such as protein similarities, expression evidence from EST alignments, SAGE or proteomics data were marked as deprecated. These genes will likely be removed in a subsequent release (and in GenBank) unless additional evidence is provided indicating they should be moved into the real gene category.

EC numbers

Enzyme Commission numbers. EC numbers constitute the numerical classification scheme for enzymes based on the chemical reactions they catalyze. EC numbers do not refer to the enzymes, but to the reactions they catalyze.

EST

Expressed sequence tag. Short (typically 100-500 base pairs) partial cDNA produced by single-shot sequencing of a cloned mRNA (cDNA) and often used to identify gene transcripts.

EST alignments

Alignments of expressed sequence tags (ESTs) with a corresponding genomic region. For example, in ToxoDB you can visualize EST alignments by clicking on the "View this sequence in the genome browser" link and turning on the EST Alignments track. Useful for identifying intron boundaries.

EST clusters

Groups of homologous, overlapping EST sequences created to reduce redundancy of the EST database.

Expression level

Level at which an mRNA or protein is present in a sample. Value can be absolute or relative to other mRNA or protein species in the sample.

Expression profile

Pattern of expression of one or more genes or proteins over time or over a set of experimental conditions (for example, during development or treatment, or as a result of a genetic mutation such as a knockout).

Expression profile correlation

Method for correlation of gene expression profiles with gene ontology (GO) annotations developed for the purpose of identifying groups of genes, pathways, and processes reacting in concert to experimental perturbations.

Expression timing

Timing of gene expression during a developmental, metabolic, regulatory, or other biological process or response.

GBrowse

Interactive genome browser developed by the Generic Model Organism Database (GMOD) project (www.gmod.org) that can be customized to show selected chromosomal features as well as display user-provided annotations.

GBrowse track

In the GBrowse viewer, a line of data that corresponds to a particular type of genomic information or feature and that is distinguished

by a particular shape or color.

GLEAN gene

Predicted gene sequence generated by GLEAN, an algorithm that integrates different sources of gene structure evidence (for example, gene model predictions, EST and protein sequence alignments to the genome, and SAGE or peptide tags) to produce a consensus gene prediction in the absence of known genes.

GO

Gene Ontology project. Collaborative project that has developed three structured, controlled vocabularies (ontologies) to describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner. The use of a consistent vocabulary allows genes from different species to be compared based on their GO annotations.

GO component

Gene Ontology term used to describe a cellular component, or the location where a gene product may act, rather than physical features of proteins or RNAs. For example, membrane (GO:0016020), extrinsic to membrane (GO:0019898), and integral to membrane (GO:0016021).

GO function

Gene Ontology term used to describe the molecular function of a gene product, the jobs that it performs, or the "abilities" that it has (for example, transporting compounds, binding to things, holding things together, and changing one thing into another). This is different from the biological processes the gene product is involved in, which involve more than one activity.

GO process

Gene Ontology term used to describe a biological process, a recognized series of events, or molecular functions associated with a gene product. A biological process is not equivalent to a pathway, though some GO terms do describe pathways.

GO term

Gene Ontology term. The building blocks of the Gene Ontology, each term is assigned to one of the three ontologies: molecular function, cellular component, or biological process. Each GO term consists of a unique alphanumerical identifier, a common name, synonyms (if applicable), and a definition. When a term has multiple meanings depending on species, the GO uses a "sensu" tag to differentiate among them. For example, the enzyme fumarase has the GO term GO:0004333, fumarate hydratase activity (fumarase activity), catalysis of the reaction: (S)-malate = fumarate + H2O.

GPI anchor

C-terminal post-translational modification of many eukaryotic proteins. The two fatty acids within the glycophosphatidylinositol (GPI) moiety anchor the protein to the outer leaflet of the plasma membrane. GPI-anchored proteins are believed to be involved in signal transduction and immune responses, as well as the pathobiology of many parasites.

Gametocyte

Eukaryotic germ cell that divides by mitosis to generate other gametocytes or by meiosis to generate gametes. Male gametocytes are called spermatocytes, and female gametocytes are called oocytes. Term often used to describe gametocytes of Plasmodium.

GenBank protein record

Protein sequence file in the GenBank database generally derived by translation of a related nucleotide record.

GenPept protein

Protein record from the GenPept database at the NCBI GenBank, which contains inferred translations of protein-coding sequences.

Gene

Fundamental physical and functional unit of heredity. Ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product, such as a protein or RNA molecule. A gene may have a number of parts, including the promoter region, untranslated regions (5' and 3' UTRs), introns, and exons.

GeneDB

Project developed by the Sanger Institute Pathogen Sequencing Unit (PSU) and aimed at developing and maintaining curated

database resources for all projects handled by the PSU. The database is accessible at www.genedb.org

Genetic markers

Known DNA sequences that can be identified by a simple assay. Generally genetic variations caused by mutation or alterations in loci that can be observed, examples include restriction length polymorphisms (RFLPs), short tandem repeats (STRs), variable number tandem repeats (VNTRs), short DNA sequences surrounding single base-pair changes (single nucleotide polymorphisms, or SNPs), or longer microsatellite sequences.

Genomic context

Location of a gene in the genome, which can influence the expression of the gene and functional interactions of the gene expression products. In this database, genes are depicted on individual gene pages with their surrounding genomic region and annotations.

Genotyped SNPs

Single nucleotide polymorphisms (SNPs) identified during genotyping of individual organism strains. See SNP genotyping.

Genotyping

Process of determining the genotype of an individual with a biological assay using PCR, DNA sequencing, or hybridization to DNA microarrays or beads. Provides a measurement of the genetic variation between members of a species.

HMM

Hidden Markov model. Statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications.

Homolog

Related by evolutionary descent, either between species (ortholog) or within a species (paralog).

Hydropathy

Hydropathicity. Degree to which a peptide or protein is likely to be soluble in water. Protein hydropathy plots can be useful in predicting transmembrane domains, potential antigenic sites, and regions that are likely to be exposed on the protein's surface.

Intergenic region

Stretch of DNA located between genes that may act to regulate gene expression.

Intersect

Similar to the Boolean operator "AND", an action used in the query history to find items that are common to two query result sets. For example, to search for items that appear in both result set X and result Y, type, "X INTERSECT Y".

Join

Similar to the Boolean operator "UNION", an action used in the query history to combine query sets. For example, to combine result sets X and Y, type, "X JOIN Y".

JBrowse

Interactive genome browser (jbrowse.org) developed by the Generic Model Organism Database (GMOD) project (www.gmod.org) that can be customized to show selected chromosomal features as well as display user-provided annotations.

KEGG map

Metabolic or regulatory interaction pathway generated by the Kyoto Encyclopedia of Genes and Genomes (KEGG) or by the use of their tools (www.genome.jp/kegg).

Locus

Position on a chromosome of a gene, feature (such as a telomere), or other chromosomal marker; also, the DNA at that position. Use of this term is sometimes restricted to mean expressed DNA regions.

Low complexity

Pertaining to sequence regions that have an unusually repetitive nature (for example, a protein sequence of low complexity might look like PPTDPPPPKKDGGPPL, and a low-complexity nucleotide sequence might be AAATAAAAAAAAATAAAAAAATTA). Low-complexity regions can create problems in sequence similarity searching by causing artifactual hits. For this reason, filters are often used to remove low-complexity sequences. Low-complexity regions also contribute to antigenic variation in apicomplexan parasites.

Mass spec data

Mass spectrometry data. Mass spectrometry is an analytical technique used to measure the mass-to-charge ratios of small molecules in several applications, including identification of proteins or peptides. In our databases, the "Identify Genes by Mass Spec Evidence" query is used to identify genes that have evidence of protein expression based on mass spec data.

Metabolic pathways

Series of chemical reactions occurring within a cell and often catalyzed by enzymes. In a pathway, a molecule is often changed or modified into another product, which can be stored by the cell, used as a metabolic product, or used to initiate another pathway.

Microarray

Microscopic array of biological molecules (for example, DNA or protein) used to determine the presence and/or amount (referred to as quantitation) of other biomolecules (other proteins, transcripts, etc.) in biological samples.

Microsatellite

Polymorphic locus in nuclear and organellar DNA that consists of repeating units of 1-4 base pairs in length. Mostly neutral and codominant, microsatellites are used as molecular markers and to study gene dosage (looking for duplications or deletions of a particular genetic region). Also known as simple sequence repeats (SSRs).

Microsatellite map

Map of microsatellite locations and linkages on a genome.

Mitochondrion

Organelle responsible for respiration in a eukaryotic cell. Proteins required for mitochondrial function are encoded both in the nucleus and within the smaller mitochondrial genome.

Motif search

Tool used to identify and locate sequence patterns (motifs) in protein and nucleic acid sequences. In our databases, this flexible search can be based on the general characteristics of the pattern and not solely on specific sequences (for example, Cys-[9-11 amino acids]-Cys or Leu-Leu-[basic residue]-Val). This allows the user to query using previously undescribed motifs.

NCBI

National Center for Biotechnology Information. Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

Nonredundant protein DB (NRDB)

Peptide sequence database containing all nonidentical protein sequences from the same species extracted from GenPept, NCBI RefSeq, Swiss-Prot, and PRF databases. Used for BLAST protein database searches because its smaller size results in shorter search times and more meaningful statistics.

Nucmer

NUCleotide MUMmer. Part of the MUMmer alignment package, an alignment tool used for the rapid alignment of very large DNA and amino acid sequences.

ORF

Amino acid sequences computed by translating the six frames of raw genomic sequence using the standard genetic code. We save the translated sequences having at least 50 amino acids. The sequences are not annotated nor human reviewed. ORF's do not necessarily begin with methionine residues, but they all terminate with a stop codons.

Oligo

Oligonucleotide. Short sequence of DNA or RNA, typically of 20-70 nucleotides.

Oligonucleotide microarray

Collection of microscopic oligonucleotide spots arrayed on a solid surface by covalent attachment to chemically suitable matrices. Used for expression profiling, the monitoring of the expression levels of hundreds or thousands of genes simultaneously. Probes for oligonucleotide microarrays are designed to match parts of known or predicted mRNAs.

Open Reading Frame

See ORF

OrthoMCL

Genome-scale algorithm for grouping orthologous protein sequences. It provides genes shared by two or more species/genomes and also genes representing species-specific gene expansion families. Therefore, it serves as a utility for automated eukaryotic genome annotation and phylogenetic profiling. Available at orthomcl.org

Ortholog

Same gene in different organisms; having evolved from the same ancestral locus.

Ortholog group

Orthologous genes shared by group of organisms; the group of genes can also contain paralogs.

Orthology-based phylogenetic profile

Tool used to find genes that are present or not present in a desired group of organisms on the tree of life (currently computed for 81 complete genomes). The user has control over the profile and over whether or not genes must be found in any particular group of organisms (for example, in Apicomplexa but not in mammals). Taxa can also be marked as indifferent (for example, it does not matter if the gene is also found in plants). Ortholog and paralog relationships are determined using the OrthoMCL algorithm.

PATS

Neural network analysis tool that identifies amino acid sequences within a query sequence that are potentially targeted to the apicoplast matrix of Plasmodium falciparum.

PDB 3D structure

Three-dimensional macromolecular structure in the Protein Data Bank (PDB) (www.pdb.org) obtained by one of three methods: X-ray crystallography (over 80%), solution nuclear magnetic resonance (NMR) (about 16%), or theoretical modeling (2%). A few structures were determined by other methods.

PROSITE motif

Protein sequence pattern or profile derived from multiple alignments of homologous sequences and stored in the PROSITE database (prosite.expasy.org), an annotated collection of motif descriptors dedicated to the identification of protein families and domains.

Paralog

Related by gene duplication within a genome; originated by duplication and then diverged from the parent copy by mutation and selection or drift.

Pearson correlation

Pearson Product Moment Correlation, the most common measure of the correlation between two variables. Reflects the degree of linear relationship between two variables and ranges from +1 to -1, with a correlation of +1 indicating a perfect positive linear relationship between variables.

Peptide mass fingerprinting

Analytical technique for protein identification wherein an unknown protein of interest is cleaved into peptides by a protease such as trypsin, and the peptides resulting from this cleavage are analyzed using a mass spectrometric method such as MALDI-TOF or ESI-

TOF. The masses derived for the peptides are then compared to a database containing known protein sequences or even to the genome. Computer programs theoretically cut the protein sequences in the database into peptides with the same protease (for example trypsin), and calculate the absolute masses of the peptides from each protein. They then compare the masses of the peptides of the unknown protein to the theoretical peptide masses of each protein encoded in the genome. The results are statistically analyzed to find the best match.

Pfam domain

Conserved protein region in the Pfam database (Pfam.org), a collection of multiple sequence alignments and hidden Markov models covering many common protein families. The alignments may represent evolutionarily conserved structures that may shed light on protein function. Profile hidden Markov models (profile HMMs) built from the Pfam alignments can be useful for associating a new protein to a known protein family, even if the homology is weak. Unlike standard pairwise alignment methods (for example, BLAST and FASTA), Pfam HMMs deal sensibly with multidomain proteins.

Phylogeny

Historical relationships among lineages of organisms or their parts, including their genes.

PlasmoAP

Algorithm/tool that predicts the likelihood that a protein sequence is targeted to the apicoplast. It provides the position of signal peptide cleavage sites in amino acid sequences if targeting is predicted.

PlasmoCyc

Database/utility built by analyzing the genomes of the Plasmodium species in EuPathDB with SRI International's pathway tools; used for searching and visualizing Plasmodium metabolic pathway information.

ProDom

Database of protein domain families generated from the global comparison of all available protein sequences (prodom.prabi.fr).

Promoter

Regulatory region of DNA located upstream (towards the 5' region) of a gene and providing a control point for regulated gene transcription.

Protein-coding

Capable of encoding a protein sequence; generally refers to a sequence of DNA.

Proteomics

The large-scale study of proteins, particularly of the full set of proteins encoded by a genome.

Pseudogene

Defunct relatives of known genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell. Although they may have some gene-like features (such as promoters, CpG islands, and splice sites), they are nonetheless considered nonfunctional due to their lack of protein-coding ability resulting from various genetic disablements (stop codons, frameshifts, or a lack of transcription) or their inability to function as an RNA (such as with rRNA pseudogenes).

PubCrawler

Free service that scans daily updates to the NCBI Medline (PubMed) and GenBank databases and alerts users to any relevant updates. Available at pubcrawler.gen.tcd.ie

Query

Sequence or term used in a database search. For example, the sequence submitted for a BLAST search is the query sequence.

RNA predictions

Predictions of genes that encode nonprotein-encoding RNA's such as tRNA, snoRNA, rRNA, etc.

RefSeq mRNA

Nonredundant mRNA sequence in the RefSeq database. RefSeq mRNA sequences with an NM_XXXXXX accession are curated sequences and are, therefore, considered more reliable than those with XM_XXXXXX accessions (predicted mRNA sequences).

Reference genome sequence

The community gold standard genome sequence. Usually the most complete assembly including an annotation. Reference genome sequences do change as superior assemblies and annotation become available.

RefSeq noncoding RNA

Nonredundant noncoding RNA (ncRNA) sequence in the RefSeq database. RefSeq ncRNA sequences with an NR_XXXXXX accession are curated sequences and are, therefore, considered more reliable than those with XR_XXXXXX accessions (predicted ncRNA sequences).

RefSeq protein

Nonredundant protein sequence in the RefSeq database. RefSeq protein sequences with NP_XXXXXX accessions are curated sequences and are, therefore, considered more reliable than those with XP_XXXXXX accessions (predicted protein sequences).

RefSeq

NCBI reference sequences. A curated nonredundant collection of sequences representing genomes, transcripts, and proteins as annotated by NCBI (available at www.ncbi.nlm.nih.gov/refseq). The annotation in these records is often different from the original GenBank submission, which may not be updated every time new information is obtained.

Repeat regions

Sequences present in many identical or highly similar copies in the genome.

SAGE tags

Serial analysis of gene expression (SAGE) tags. Short (14-nucleotide) sequences found within mRNA, the relative abundance of which indicates the level of expression of the mRNA containing that tag.

SNP

Single nucleotide polymorphism. Small genetic changes or variations that can occur within a DNA sequence, for example when a single nucleotide, such as an A replacing one of the other three nucleotide letters C, G, or T. Most SNPs are found outside of coding sequences, but SNPs found within a coding sequence are more likely to alter the biological function of a protein. SNPs may be synonymous (generating a conservative change not altering the amino acid sequence) or they can be nonsynonymous and change the amino acid that is encoded.

SNP density

Amount or number of single nucleotide polymorphisms (SNPs) in a region of the genome.

SNP genotyping

Identifying and mapping single nucleotide polymorphisms (SNPs) in an effort to determine the genotype members of a species. SNPs usually consist of two alleles (where the rare allele frequency is less than 1%), are evolutionarily conserved, and are the most common type of genetic variation. See Genotyping.

Scaffolds

In genomic mapping, a series of contigs that are in the right order and orientation, but not necessarily connected in one continuous stretch of sequence.

Sequence similarity

Degree of similarity between two or more protein or nucleotide sequences.

Signal peptide

Short (3-60 amino acids) peptide sequence that directs the co-translational import of a protein to certain organelles or for secretion.

SignalP

Program that predicts the presence and location of signal peptide cleavage sites in amino acid sequences from Gram-positive and -negative prokaryotes and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/nonsignal peptide prediction based on a combination of several artificial neural networks and hidden Markov models (HMMs).

Strand

One half of the DNA helix; string or stretch of covalently linked nucleotides.

Subtract

Similar to the Boolean operator "NOT", an action used in the query history to remove items from one result set that occur in another result set. For example, to remove items that exist in set Y from those in set X, type, "X SUBTRACT Y".

Syntenic

Loci located on the same chromosome but not necessarily linked. For example, genes that are part of a syntenic group share a common chromosomal location. Also used to refer to conservation of gene order across species.

TIGR

The Institute for Genomic Research, a nonprofit center dedicated to deciphering and analyzing genomes.

TM domains

See transmembrane domain.

TWINSCAN gene models

Gene models generated using TWINSCAN, a tool that integrates traditional probability models (such as those underlying GENSCAN and FGENESH) with information from alignments between two genomes. TWINSCAN is based on the idea that functional sequences show different patterns of evolutionary conservation than sequences under little selective pressure, such as the central regions of introns. TWINSCAN is designed for the analysis of high-throughput genomic sequences containing an unknown number of genes.

Telomere

Nucleoprotein complexes that constitute the physical ends of linear eukaryotic chromosomes and that have important functions, primarily in the protection, replication, and stabilization of the chromosome ends. Telomeres often contain lengthy stretches of tandemly repeated simple DNA sequences composed of a G-rich strand and a C-rich strand (called terminal repeats). These terminal repeats are highly conserved. Sequences adjacent to the telomeric repeats are often highly polymorphic and rich in repetitive elements (termed subtelomeric repeats); in some cases, genes have been found in the proterminal regions of chromosomes.

TigrScan gene

Gene model generated using TigrScan, a gene-finding tool based on the generalized hidden Markov model (HMM) framework, similar to GENSCAN and Genie. It is highly reconfigurable and includes software for retraining.

ToxoCyc

Database/utility built by analyzing the Toxoplasma gondii genome with SRI International's pathway tools; used for searching and visualizing Toxoplasma metabolic pathway information.

Translation

Synthesis of protein from an mRNA template.

Transmembrane domain

Three-dimensional protein structure that is thermodynamically stable in a membrane. This may be a single alpha helix, a stable complex of several transmembrane alpha helices, a transmembrane beta barrel, a beta-helix of gramicidin A, or any other structure. Transmembrane domains average 20 amino acid residues in length, though they may be much smaller or much longer.

Transmembrane protein

Protein that spans an entire biological membrane.

UTR

Untranslated region. Section of messenger RNA (mRNA) that either precedes (5' UTR) or follows (3' UTR) the coding region and is not itself translated. The UTR contains several regulatory regions, including the polyadenylation (polyA) site in the 3' UTR, sequences involved in the initiation of translation (in the 5' UTR), and binding regions for proteins and other regulatory molecules in both the 3' and 5' UTR.

UniGene

Project and database at NCBI aimed at defining gene-oriented clusters of expressed sequence tags (ESTs). Sets of ESTs are clustered based on strong sequence homology in an attempt to define a specific, nonredundant cluster for each transcript in a tissue or genome. Each UniGene cluster contains sequences that represent a unique gene in addition to information about the tissue types in which the gene has been expressed and map location.

Wildcard character

Character used to substitute for any other character(s) in a string.

Xenolog

Gene found in an unrelated species and that is related by gene transfer rather than common vertical descent.

blastn

Version of the basic local alignment search tool (BLAST) used to compare a nucleotide query sequence against a nucleotide sequence database.

blastp

Version of the basic local alignment search tool (BLAST) used to compare a protein query sequence against a protein sequence database.

blastx

Version of the basic local alignment search tool (BLAST) used to compare a nucleotide query sequence translated in all reading frames against a protein sequence database.

cDNA

Complementary DNA. DNA molecule synthesized by the enzyme reverse transcriptase using an mRNA as template.

cDNA microarray

Collection of microscopic cDNA spots commonly representing single genes and arrayed on a solid surface (commonly glass slides) by covalent attachment to chemically suitable matrices. Used for expression profiling, the monitoring of the expression levels of hundreds or thousands of genes simultaneously.

ePCR

Electronic PCR (polymerase chain reaction). Computational procedure used to check for uniqueness in spacing and number of primer binding sites within DNA sequences. Searches for subsequences that closely match a set of PCR primers and that have the correct order, orientation, and spacing to make a PCR product. Used to check the expected length of a PCR product, which can provide information regarding unexpected repetitive sequences.

ncRNA

Noncoding RNA. Any RNA that is not translated into a protein. Includes transfer RNA (tRNA), ribosomal RNA (rRNA), small RNAs such as snoRNAs, microRNAs, siRNAs and piRNAs, as well as long ncRNAs.

rRNA

Ribosomal RNA. Component of the ribosomes, which function in protein synthesis.

snRNA

Small nuclear RNA. Class of small RNA molecules found within the nucleus, transcribed by RNA polymerase II or III, and involved in a variety of important processes such as RNA splicing (removal of introns from hnRNA), regulation of transcription factors (7SK RNA) or RNA polymerase II (B2 RNA), and maintaining telomeres. They are always associated with specific proteins, and the complexes are referred to as small nuclear ribonucleoproteins (snRNP) or snurps. These elements are rich in uridine. A large group of

snRNAs known as small nucleolar RNAs (snoRNAs) are small RNA molecules that play an essential role in RNA biogenesis and chemical modification of ribosomal RNAs (rRNAs) and other RNA genes (tRNA and snRNAs). They are located in the nucleus and the cajal bodies of eukaryotic cells (the major sites of RNA synthesis).

tRNA

Transfer RNA. Small RNA chain (73-93 nucleotides) that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. A three-base region, the anticodon, pairs to the corresponding three-base codon region on the template mRNA. Each type of tRNA molecule can be attached to only one type of amino acid, but because the genetic code contains multiple codons that specify the same amino acid, tRNA molecules bearing different anticodons may also carry the same amino acid.

tblastn

Version of the basic local alignment search tool (BLAST) used to compare a protein query sequence against a translated nucleotide sequence database.

tblastx

Version of the basic local alignment search tool (BLAST) used to compare the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.