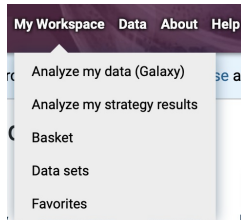


Use pre-configured workflows in VEuPathDB's Galaxy to analyze your data. Workflows include:

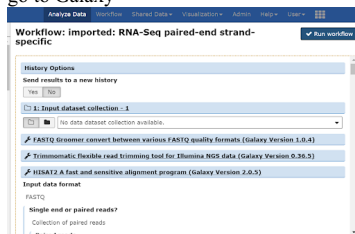
- RNA-Seq coverage across the genome and RPKM for each gene
- RNA-Seq identification of genes with statistically significant expression differences between two samples
- Assign your proteins to OrthoMCL groups
- Variant calling

For RNA-Seq analyses you'll be able to export your results to your own private dataset section in VEuPathDB. You can:

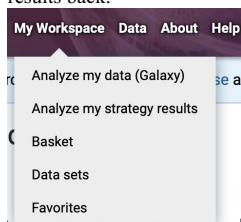
- run searches against them
- see them as columns in search results
- see them as graphs in gene pages
- see them as genome browser tracks.



Start by clicking *Analyze my data (Galaxy)* in the "My Workspace" menu to go to Galaxy



Run workflows in Galaxy, and export the results back.

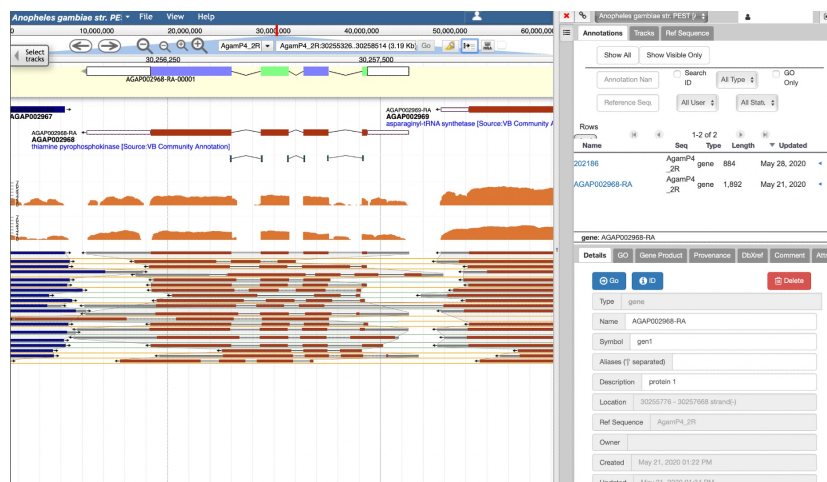


Then click on *data sets* in the "My Workspace" menu to see your data here

VEuPathDB supports curation and annotation of genomes and provides various tools for individual and community-driven projects:

1 Structural and functional annotation in Apollo.

Apollo is a collaborative, real-time, genome annotation web-based editor. It allows structural improvement of gene models via individual or collaborative efforts while taking advantage of VEuPathDB-integrated data or user-generated tracks.



2 User comments.

Functional annotation of genes can be updated via user comments. User comments are linked to user accounts and become visible immediately after submission. User comments can include files (e.g. protein location images or knockout library phenotype descriptions) or a list of relevant PubMed IDs.

NCU06658 SET domain-containing protein

Type: protein coding

Chromosome: V

Location: CM002240:4,152,430..4,154,510(-)

Species: Neurospora crassa

Strain: OR74A

Status: Reference Strain

View 1 user comment, or add a comment

Shortcuts

Synteny

BLAT Alignments

SNPs

Transcriptomics

Protein Features

Also see NCU06658 in the Genome Browser or Protein Browser

Manual curation (functional annotation).

VEuPathDB regularly curates and updates functional attributes of genes for the group of **selected genomes**. For example, gene names, gene synonyms, product descriptions, Gene Ontology annotations, EC numbers and PubMed IDs are integrated regularly from user comments and scientific publications.

GO Terms

Download

Data sets

Search this table...

Ontology	GO ID	Is Not	GO Term Name	Source	Evidence Code	Reference
Molecular Function	GO:0005515	N/A	protein binding	Interpro	IEA	N/A
Molecular Function	GO:0016279	N/A	protein-lysine N-methyltransferase activity	GeneDB	ISM	GO_REF:0000002

There are numerous ways to download raw data files from VEuPathDB.

- 1 Use [Data Downloads](#) to download current and archived genome .fasta, .gff, .gaf, etc files
- 2 You can download most tables on our gene pages

My Strategies

Searches

Tools

My Workspace

Data

About

Help

Transcripts

Download

Data sets

Search this table...

AGAP000080

expand all | collapse all

Search section names...

1 Gene models

2 Annotation, curation and identifiers

3 Link outs

Transcript	# exons	Transcript length	Protein
AGAP000080-RA	6	2739	912
AGAP000080-RB	6	2133	710

- 3 From the strategy system, add Columns of data (such as expression values) you wish to access then Download the gene results table.

Gene Results

Genome View

Analyze Results

Genes: 5,245

Transcripts: 5,254

Show Only One Transcript Per Gene

1

2

3

...

263

Rows per page: 20

Download

Add to Basket

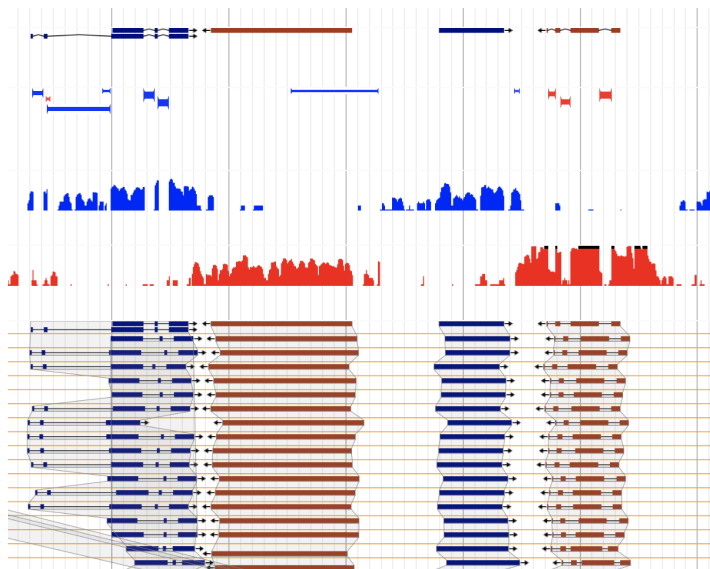
Add Columns

Gene ID	Transcript ID	WT_treated_with_sulfadiazine (PbANKA WT and AP2-G2 KO Marray)	AP2-G2_treated_with_sulfadiazine (PbANKA WT and AP2-G2 KO Marray)
PBANKA_0000101	PBANKA_0000101.1	0.18	0.33
PBANKA_0000201	PBANKA_0000201.1	N/A	N/A
PBANKA_0000301	PBANKA_0000301.1	0.8	0.08

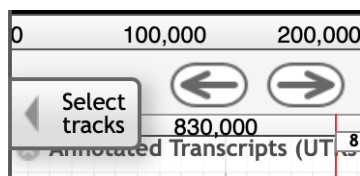
- 4 You can utilize our [Web Services](#) to programmatically access our searches.

Please go to a component site to access JBrowse.

VEuPathDB provides genome browsing capability through the JBrowse genome browser. JBrowse may be accessed directly from the tools section located in the menu bar in the header or from relevant sections on gene pages of interest. JBrowse provides numerous functionalities including:



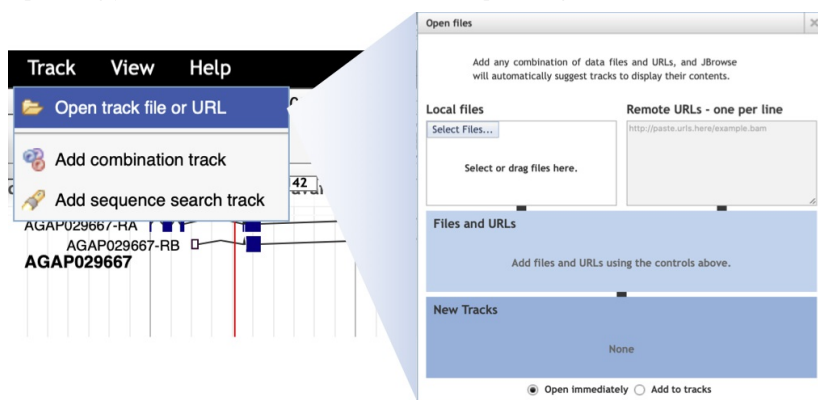
- 1 Data tracks that can be turned on or off from the select tracks tab.



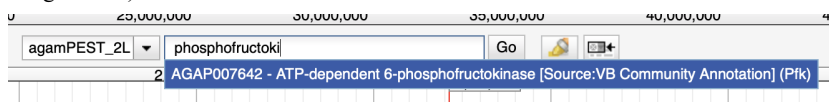
- 2 Zooming and panning across your genome of interest.



- 3 Uploading your own data tracks, such as RNA-Seq coverage data.



- 4 Search capabilities both with key words (screen shot below) and regular expressions (select the add sequence search track shown in image above).



VEuPathDB is packed with data, tools and visualizations that can help answer your research questions. We gather data from many sources, analyze according to standard workflows, and present the results for you to mine in a point and click interface. Here's how to get started:

SITE SEARCH: Explore the site; find what you need

Enter a term or ID in the site search box at the top of any page. The site search finds documents and records that contain your term and returns a summary of categorized matches. Its easy to find genes, pathways, searches, data sets and more with the site search.

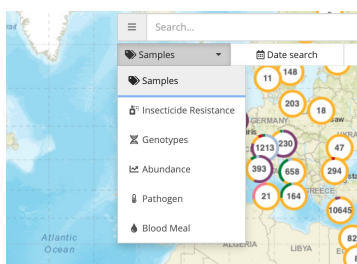
SEARCH STRATEGY SYSTEM: Discover meaningful biological relationships from large volumes of data

The strategy system offers over 100 structured searches that can be combined to produce multi-step strategies. Each search queries a specific data set and returns a list of records that share the biological characteristic defined by the data. To run a search, navigate to one from the Search For... menu on the left, adjust the search parameters if desired, and then click Get Answer. Your results are presented in the My Strategies section.

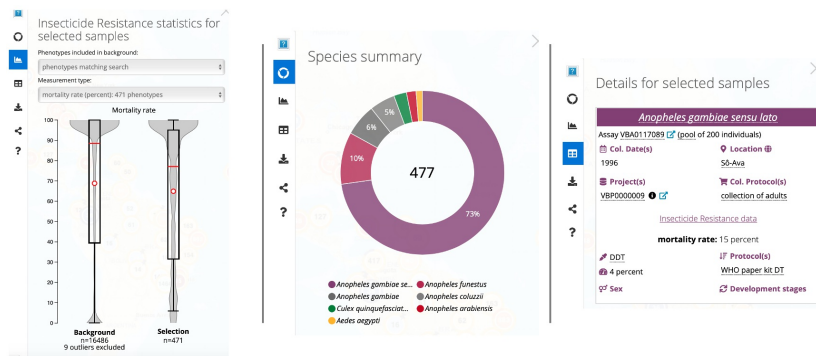
MapVEu

[MapVEu](#) integrates genomic, phenotypic and population data for traits such as insecticide resistance genotypes and phenotypes, genetic variation with microsatellites, chromosomal inversions and SNPs, population abundance, pathogen infection status and blood meal identification.

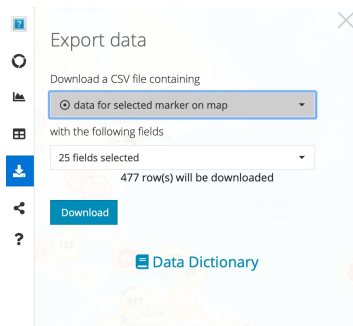
1 Six available map views.



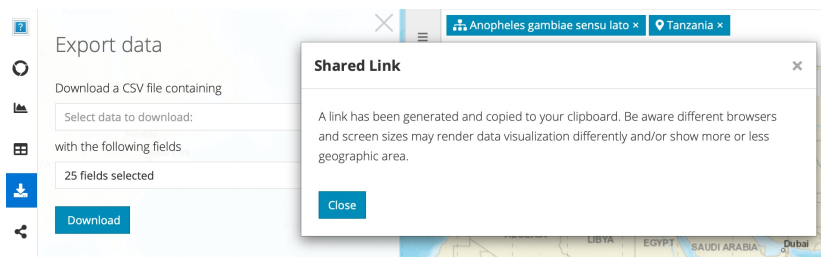
2 Data is displayed as circular markers that are responsive to zoom level. Clicking on a marker provides access to rich metadata and plotting options.



- Raw data is available for download and export.



- Search results may also be shared.



Phenotypic evidence includes high-throughput phenotype data from experiments involving RNAi target sequencing, whole-genome knockout collections, piggyBac insertion mutagenesis, and also manually curated phenotypes by VEuPathDB and other resources.



Phenotypic data is available to you as:

- Tables (text and image data) associated with various phenotype collections on individual gene record pages.

8 Phenotype

Neurospora Genome Project Phenotype Image Collection Download Data sets

Search this table...

Temperature	Location	Timepoint	Media	Morphology	Physiology	Sexual Development	Image
25C	edge	24hr	minimal	Normal growth pattern (Plates) normal aerial hyphae formation (Slants) normal conidiation (Slants) normal pigmentation normal pigmentation (Plates)	aerial hyphae extension linear growth	N/A	
25C	edge	24hr	minimal	Normal growth pattern (Plates) normal aerial hyphae formation (Slants) normal conidiation (Slants) normal pigmentation normal pigmentation (Plates)	aerial hyphae extension	N/A	

Phenotypic analysis of Neurospora crassa knockout mutants Download Data sets

Mating Type	Basal Hyphae Growth Rate	Aerial Hyphae Height	Conidia Number	Conidia Morphology	Protoperithecia Number	Protoperithecia Morphology	Perithecia Number	Perithecia Morphology	Ascospore Number	Ascospore Morphology
mata	82.5	31	Normal	Normal	Normal	Normal	Normal	Normal	Reduced	Normal

- Searches that query datasets for genes based on their observable characteristics of genetic manipulations, outcomes of host-pathogen interactions, CRISPR and quantitative phenotypes, etc.

Search for...

expand all | collapse all

Filter the searches below...

- Genes
 - Annotation, curation and identifiers
 - Function prediction
 - Gene models
 - Genetic variation
 - Genomic Location
 - Immunology
 - Orthology and synteny
 - Pathways and interactions
 - Phenotype
 - Phenotype Evidence

Identify Genes based on Phenotype Evidence

Filter Data Sets:

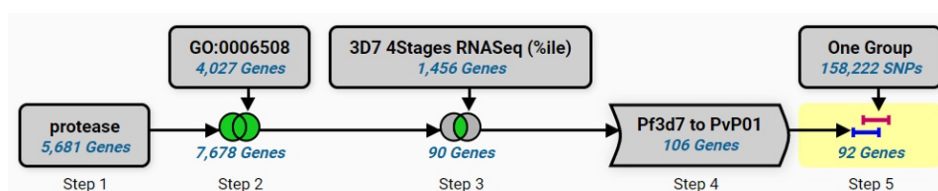
Legend: ☒ Curated Phenotype ☐ Phenotype Test

A. fumigatus AZ93	Manually Curated <-Aspergillus>- Phenotypes (Eukaryotic Pathogen Database Resources (EuPathDB))	<input type="button" value="CP"/>
A. nidulans F502_A4		
A. niger CBS 513.18		
A. oryzae H940		
C. parv WM276	Manually Curated <-Cryptococcus>- Phenotypes (Eukaryotic Pathogen Database Resources (EuPathDB))	<input type="button" value="CP"/>
C. neoformans var. grubii H99		
C. neoformans var. neoformans JEC21		
F. graminearum Pf-1	Manually Curated <-Fusarium>- Phenotypes (Eukaryotic Pathogen Database Resources (EuPathDB))	<input type="button" value="CP"/> <input type="button" value="PT"/>
N. crassa OR74A	Phenotypic analysis of Neurospora crassa knockout mutants (Borkovich et al.)	<input type="button" value="CP"/>
N. crassa OR74A	Neurospora Genome Project Phenotype Image Collection (Dunlap et al.)	<input type="button" value="CP"/>

3 Columns in the search results table.

Gene ID	Transcript ID	Organism	Product Description	Fold Change	High-throughput phenotyping using RNAi target sequencing - fpkm Graph
Tb927.9.16210	Tb927.9.16210;pseudogenic_transcript	T. brucei brucei TREU927	expression site-associated gene 3 (ESAG3, pseudogene), putative	-362.8*	Tb927.9.16210 - Unique CDS aligned

All Available Searches



Discover meaningful biological relationships from large volumes of data with VEuPathDB's unique and powerful **Search Strategies** - a tool for mining Omics data without programming experience.

1 We analyze the data.



2 You search our analysis results from our point and click interface. Choose from >100 searches that query ~20 data types and ~1000 data sets. Access the searches from the homepage 'Search for...' panel or from the dropdown Searches menu in every page header.

Search for...

Genes

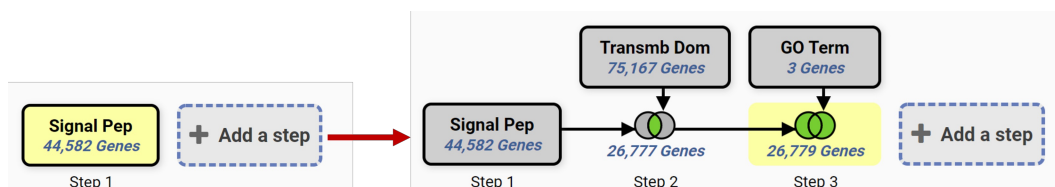
- Protein targeting and localization
 - Exported Protein
 - Pf. Subcellular Localization
 - Predicted Signal Peptide
 - Transmembrane Domain Count

Search for genes with predicted signal peptides

Signal Pep
44,582 Genes

Step 1

3 Add searches one at a time to create a strategy and find exactly the type of genes you need, such as signal peptide containing genes that are expressed in merozoites.



VEuPathDB integrates unpublished, pre-publication, and published genomic-scale datasets associated with the diverse eukaryotic microbes, fungi, vectors of human disease, and hosts.

While some data can be accepted directly from the provider (for example, proteomics data in Excel spreadsheets or population data with GPS coordinates), all sequence data such as genomes, RNA-Seq or ChIP-Seq data sets, must be available from an INSDC database (DDBJ, ENA, or GenBank).

Plan ahead and let us know of the data availability in advance to reserve your spot in the submission queue. We have data releases every two months and will get in contact with you once your data is scheduled for the release.

4 [Read more on VEuPathDB data submission and release policies](#)

2 [Nominate a data set for integration into VEuPathDB](#)

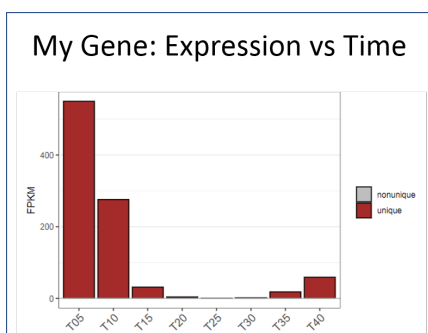


VEuPathDB Dataset and Bioinformatics Tool Nomination Form

[All RNA-Seq Searches](#) [Example Gene Page Data](#) [All Microarray Searches](#)

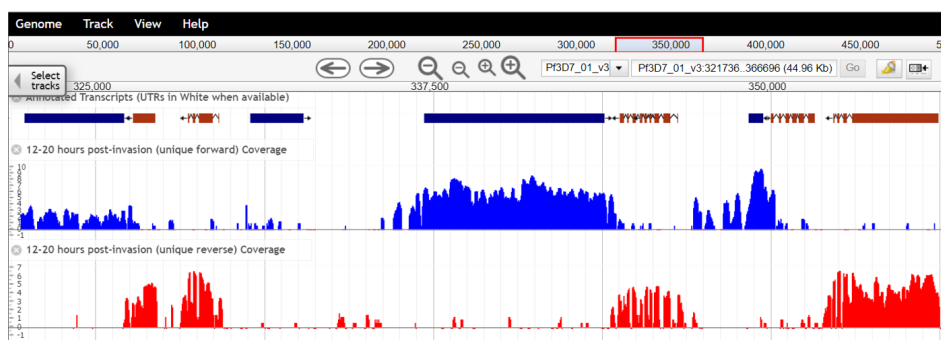
VEuPathDB supports transcriptomic data mining from RNA sequence, microarray, and expressed sequence tags. We analyze pertinent data from your research community and make our analysis results available to you as:

1 Tables, and graphs on gene pages. [Example Gene Page Data](#)



Gene	Sample	Value	Percentile 1
PF3D7_1133400	T05- sense - unique	549.82	96.9
PF3D7_1133400	T10- sense - unique	275.92	92.32
PF3D7_1133400	T15- sense - unique	31.51	64.51
PF3D7_1133400	T20- sense - unique	4.03	38.45
PF3D7_1133400	T25- sense - unique	0.75	15.1
PF3D7_1133400	T30- sense - unique	1.69	15.8
PF3D7_1133400	T35- sense - unique	18.22	34.6
PF3D7_1133400	T40- sense - unique	58.96	66.89

2 Genome browser tracks for dynamic visualization of analyzed data as read coverage plots or predicted introns. Sequence data is mapped to a reference genome and made available in JBrowse. Scroll, zoom or stack data for more data mining power. [Example JBrowse view](#)



3 Searches that query individual [RNA Seq](#) or [Microarray](#) data sets and return a list of genes based on sample characteristics such as life cycle stage or treatment. Associated gene data can be added to the search result via the Add Columns tool. We offer several query types including:

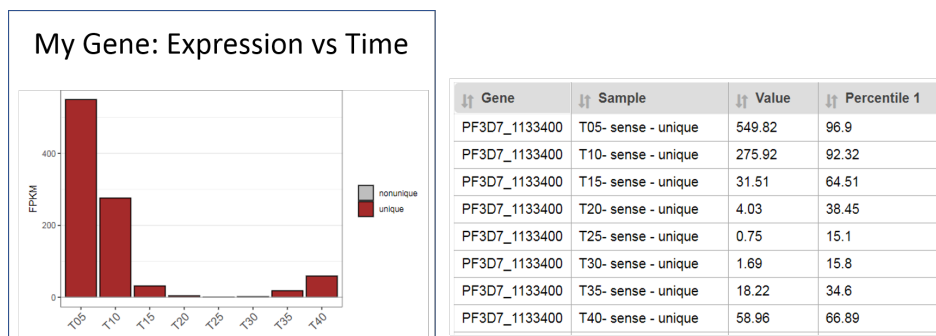
- differential expression (Fischer's exact test),
- fold change (algebraic calculation),
- percentile rank within a sample,
- sense/antisense comparisons, and

- expression profile similarity.

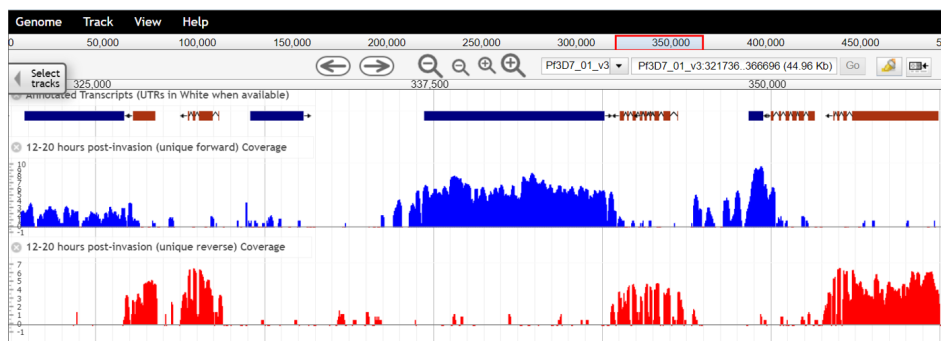
In addition, [My Workspace](#) offers a private Galaxy workspace for analyzing your own data and porting the results to VEuPathDB to compare with public data. All VEuPathDB genomes are pre-loaded and several pre-configured workflows for RNA-seq analysis and variant calling are available. All RNA-Seq Searches

VEuPathDB supports transcriptomic data mining from RNA sequence, microarray, and expressed sequence tags. We analyze pertinent data from your research community and make our analysis results available to you as:

- 1 Tables and graphs on gene pages.



- 2 Genome browser tracks for dynamic visualization of analyzed data as read coverage plots or predicted introns. Sequence data is mapped to a reference genome and made available in JBrowse. Scroll, zoom or stack data for more data mining power.



- 3 Searches that search individual RNA-Seq or Microarray data sets and return a list of genes based on sample characteristics such as life cycle stage or treatment. Associated gene data can be added to the search result via the Add Columns tool. We offer several search types including:

- differential expression (DEseq2),
- fold change (algebraic calculation),
- percentile rank within a sample,
- sense/antisense comparisons, and
- expression profile similarity.

In addition, [My Workspace](#) offers a private Galaxy workspace for analyzing your own data and porting the results to VEuPathDB to compare with public data. All VEuPathDB genomes are pre-loaded and several pre-configured workflows for RNA-Seq analysis and variant calling are available.

Welcome to the new VectorBase *beta* website!

Welcome! VectorBase *beta* has a new look and new tools to help you better browse and Search vector omics and population data. Feel free to explore the other cards in the **Overview of Resources and Tools** section to learn more.

The [legacy vectorbase.org](#) site is still available to help you transition to the new site, if needed.

To view a summary of the datasets available for each species, follow this link in the header menu [Data => Organisms - Data type summary](#)

We will start to incorporate new data sets in the second half of 2020. Please let us know if you wish to submit a data set or make us aware of one relevant to your research, to help us prioritize our list.

If you have questions, concerns, suggestions or need additional help please [contact us](#). Specific feedback about the new site is greatly appreciated by completing one of these surveys:

- [Home page survey](#)
- [Site search survey](#)
- [Search strategy survey](#)

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) - see [funding section below](#). 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.
- [PATRIC](#) & [VIPR](#) focus on bacterial and viral pathogens.

In addition to support from NIAID, VEuPathDB 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.

- [Gene Metrics Table](#) summarizes the number of annotated genes for the organisms currently supported by this website.
- [Genome Summary](#) 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](#).

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:

À Á À	À Á À AmoebaDB :À Á À	À Á À MicrosporidiaDB :À Á À	À Á À ToxoDB :À Á À	À Á À
À Á À	À Á À William PetriÀ Á À	À Á À Gira BhabhaÀ Á À	À Á À David RoosÀ Á À	À Á À
À Á À	À Á À Serge AnkriÀ Á À	À Á À Pattana Jaroenlak (Michael)À Á À	À Á À TrichDB :À Á À	À Á À
À Á À	À Á À Craig RobertsÀ Á À	À Á À Damian EkiertÀ Á À	À Á À Antonio Pereira-NevesÀ Á À	À Á À
À Á À	À Á À Fiona HenriquezÀ Á À	À Á À Michael CammerÀ Á À	À Á À Marlene BenchimolÀ Á À	À Á À
À Á À	À Á À Hugo Aguilar-DÃ-azÀ Á À	À Á À PiroplasmaDB :À Á À	À Á À TriTrypDB :À Á À	À Á À
À Á À	À Á À Julio C. CarreroÀ Á À	À Á À Ellen YehÀ Á À	À Á À Rick TarletonÀ Á À	À Á À
À Á À	À Á À CryptoDB :À Á À	À Á À Lowell KappemeyerÀ Á À	À Á À Richard WheelerÀ Á À	À Á À
À Á À	À Á À Boris StriepenÀ Á À	À Á À Audrey LauÀ Á À	À Á À Leandro Lemgruber SoaresÀ Á À	À Á À
À Á À	À Á À FungiDB :À Á À	À Á À Dirk DobbelaereÀ Á À	À Á À Margaret MullinÀ Á À	À Á À
À Á À	À Á À ZyGoLife À Á À	À Á À Manoj DuraisinghÀ Á À	À Á À Camila Silva GonÃalvesÀ Á À	À Á À
À Á À	À Á À Jason StajichÀ Á À	À Á À Brendan ElsworthÀ Á À	À Á À Wanderley de SouzaÀ Á À	À Á À
À Á À	À Á À Zachary LewisÀ Á À	À Á À Caroline KeroackÀ Á À	À Á À Maria Cristina Machado MottaÀ Á À	À Á À
À Á À	À Á À GiardiaDB :À Á À	À Á À Isabelle CoppensÀ Á À	À Á À	À
À Á À	À Á À Fran GillinÀ Á À	À Á À PlasmoDB :À Á À	À Á À	À
À Á À	À Á À Tineke LauwaetÀ Á À	À Á À Lawrence BannisterÀ Á À	À Á À	À
À Á À	À Á À Barbara DavidsÀ Á À	À Á À Lewis TilneyÀ Á À	À Á À	À

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 Papadopolou, Laurie Read, Sergio Schenkman, Rick Tarleton, Brent Weatherly, Bill Wickstead, Michael (Mick) Urbaniak

Vectors: Gregory Dasch, Jeff Grabowski, MarÅa de Lourdes MuÅoz, 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 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 MuÃoz, 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.

Ã	Ã	Ã	Ã	Ã	Ã
Ã Ã Ã	Ã Ã Ã Lyric BartholomayÃ Ã Ã	Ã Ã Ã Michael GottliebÃ Ã Ã	Ã Ã Ã Malcolm McConvilleÃ Ã Ã	Ã Ã Ã John TaylorÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã Matt BerrimanÃ Ã Ã	Ã Ã Ã Keith GullÃ Ã Ã	Ã Ã Ã Nicola MulderÃ Ã Ã	Ã Ã Ã Jake TuÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã Bill BlackÃ Ã Ã	Ã Ã Ã Matthew HahnÃ Ã Ã	Ã Ã Ã Ull MunderlohÃ Ã Ã	Ã Ã Ã Brett TylerÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã John BoothroydÃ Ã Ã	Ã Ã Ã Adrian HehlÃ Ã Ã	Ã Ã Ã Daniel NeafseyÃ Ã Ã	Ã Ã Ã Kenneth VernickÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã Greg BuckÃ Ã Ã	Ã Ã Ã Steve HiggsÃ Ã Ã	Ã Ã Ã Kenneth OlsonÃ Ã Ã	Ã Ã Ã Sarah VolkmanÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã Geraldine ButlerÃ Ã Ã	Ã Ã Ã Catherine HillÃ Ã Ã	Ã Ã Ã Bill PetriÃ Ã Ã	Ã Ã Ã Jonathan WastlingÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã Angela CruzÃ Ã Ã	Ã Ã Ã Marcelo Jacobs-LorenaÃ Ã Ã	Ã Ã Ã Barry PittendrighÃ Ã Ã	Ã Ã Ã Scott WeaverÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã George DimopoulosÃ Ã Ã	Ã Ã Ã Anthony (Tony) JamesÃ Ã Ã	Ã Ã Ã Jeffrey PowellÃ Ã Ã	Ã Ã Ã Louis WeissÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã Martin DonnellyÃ Ã Ã	Ã Ã Ã Pedro Lagerblad de OliveiraÃ Ã Ã	Ã Ã Ã Hillary RansonÃ Ã Ã	Ã Ã Ã Dyann WirthÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã Patrick DuffyÃ Ã Ã	Ã Ã Ã Greg LanzaroÃ Ã Ã	Ã Ã Ã Alexander RaikhelÃ Ã Ã	Ã Ã Ã Jennifer WortmanÃ Ã Ã	Ã Ã Ã
Ã Ã Ã	Ã Ã Ã Pascale GaudetÃ Ã Ã	Ã Ã Ã Daniel MasigaÃ Ã Ã	Ã Ã Ã Lincoln SteinÃ Ã Ã	Ã Ã Ã Guiyun YanÃ Ã Ã	Ã Ã Ã

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*, *Microsporidia* (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**

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

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 glycosphosphatidylinositol (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 AAATAAAAAAAAAAATAAAAAAATTA). 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).

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](#).

TWINSKAN gene models

[Gene](#) models generated using TWINSKAN, a tool that integrates traditional probability models (such as those underlying GENSCAN and FGENESH) with information from alignments between two genomes. TWINSKAN 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. TWINSKAN 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.

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 [Genomics Unified Schema \(GUS\) 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.

► >>> [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.

Learning about VEuPathDB

Virtual Event

Host us for your lab meeting or workshop!

Webinars

Live and on-demand instructional videos.

Workshops

In-person training to improve your skills.

Learning Materials

Explore quickstarts and tutorials.

Methods

Understand our data and its provenance.

Table 1: Core database analysis pipelines and hive components

Data	Program	Parameters	Version
EST alignments	BLAT	-nohead -maxIntron=1000 -t=dna -q=dna -dots=10	35
Open Reading Frames	orfFinder	--minPepLength 50	https://github.com/VEuPathDB/.../orfFinder
Synteny	Mercator and MAVID	-p (PATH TO MERCATOR DIRECTORY) -t (tree string) -m (MAVID_EXE) -c (CNDSRC_DIR) -d draftGenome1... -d draftGemoneN -n nonDraftGeome1... -n nonDraftGenomeN -r referenceGenome	https://github.com/VEuPathDB/.../runMercator

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 author'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.

Genome analyses

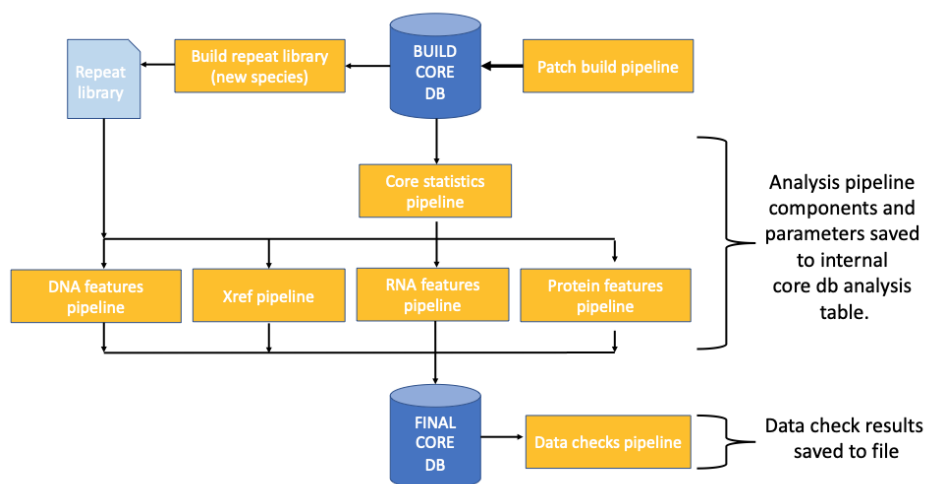
Genome sequence and annotation are analyzed by the [EBI Pipeline](#) supplemented with [three in-house analyses](#). In the rare case that the EBI pipeline cannot be applied to a genome, we use a [series of in-house analyses](#) in lieu of the EBI Pipelines.

EBI Pipeline

VEuPathDB employs the [Ensembl genome analysis](#) pipelines 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 alignments 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](#)

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. Further details coming soon.

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
Mark Caddick, 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 Amos⁴, Rachel Ankirskiy¹, Cristina Aurrecochea¹, Matthieu Barba⁹, Ana Barreto³, Evelina Basenko⁴, Wojtek Bazant², Dan Beiting², Bob Belnap¹, Ulrike Böhme⁵, John Brestelli³, Brian Brunk², Mark Caddick⁴, Danielle Callan², Mikkel Christensen⁹, George Christophides⁸, Kathryn Crouch⁶, Katie Cybulski⁷, Elaine Daugan⁴, Jeremy DeBarry¹, Ryan Doherty³, Yikun Duan², Dave Falke¹, Steve Fischer³, Paul Flicek⁹, Bindu Gajria², Gloria I. Giraldo-CalderÃ³n⁷, Omar S. Harb², Elizabeth Harper², Danica Helb², Mark Hickman², Connor Howington⁷, Sufen Hu², Jay Humphrey¹, John Iodice³, John Judkins², Sarah Kelly⁸, Jessica C. Kissinger¹, Dae Kun Kwon⁷, Kris Lamoureux¹, Daniel Lawson⁸, Wei Li², Brianna Lindsay², Jamie Long², Bob MacCallum⁸, Gareth Maslen⁹, Mary Ann McDowell⁷, Greg Milewski², Jarek Nabrzyski⁷, David S. Roos², Samuel Rund⁷, Steph Wever Schulman⁷, Achchuthan Shanmugasundram⁴, Vasili Sitnik⁹, Drew Spruill¹, David Starns⁴, Christian J. Stoeckert Jr.³, Sheena Shah Tomko², Haiming Wang¹, Susanne Warrenfeltz¹, Robert Wieck⁷, Mariann Winkelman², Lin Xu², Jie Zheng³.

Previous VEuPathDB Team members, 2004-2020

Antelmo Aguilar⁷, James Allen⁹, Alexis Allot⁹, Nora Besansky⁷, Austin Billings², Sanjay Boddu⁹, Steve Bogor⁷, Ewan Birney⁹, Andrew Brockman⁸, Robert Bruggner⁷, Ja'Shon Cade³, David Campbell⁷, Cristian Cocos², Frank Collins (VectorBase Principal Investigator, 2004-2018)⁷, Kathy Couch¹, Greg Davies⁷, Ale Diaz Miranda², Emmanuel Dialynas, Jennifer Dommer³, Vicky Dritsou, Scott Emrich¹⁰, Xin Gao², William Gelbart¹², Sandra Gesing⁷, Alan Gingle¹, Greg Grant³, Matt Guidry¹, Martin Hammond⁹, Mark Heiges¹, Christiane Hertz-Fowler (Principal Investigator, WT 2008-2019)⁴, Nicholas Ho⁸, Daniel Hughes⁹, Frank Innamorato³, San James¹⁴, Amie Jaye⁸, Fotis Kafatos⁸, Paul Kersey⁹, Ioannis Kimitzoglou⁸, Nathan Konopinski⁷, Carolyn Knoll², Eileen T. Kraemer¹, Nick Langridge⁹, Cris Lawrence², Neil Lobo⁷, Christos (Kitsos) Louis¹¹, Ross Madden⁶, Greg Madey⁷, Elisabetta Manduchi³, Karine Megy⁹, John A. Miller⁶, Elvira Mitraka¹¹, Vishal Nayak³, Cary Pennington¹, Deborah F. Pinney³, Brian Pitts¹, Jane A. Pulman⁴, Caleb Reinking⁷, Seth Redmon, Chris Ross¹, Andrew Sheehan⁷, Fatima Silva⁴, Ganesh Srinivasamoorthy¹, Scott Szakonyi⁷, Pantelis Toplais¹¹, Ryan Thibodeau¹, Charles Treatman², Betsy Wenthe¹, Matt Vander Werf⁷, Maggie Werner-Washburne¹³, Patricia L. Whetzel³, Derek Wilson⁹, Andrew Yates⁹

¹University of Georgia, Athens, GA 30602, USA

²University of Pennsylvania, Philadelphia, PA 19104, USA

³University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

⁴University of Liverpool, UK

⁵Wellcome Sanger Institute, Hinxton, CB10 1RQ, UK

⁶Wellcome Centre for Integrative Parasitology, University of Glasgow, UK

⁷University of Notre Dame, Notre Dame, IN 46556, USA

⁸Imperial College London, South Kensington, London SW7 2BU, UK

⁹European Bioinformatics Institute, Hinxton, CB10 1SD, UK

¹⁰University of Tennessee, Knoxville, TN 37996, USA

¹¹Institute of Molecular Biology and Biotechnology-FORTH, Heraklion, Crete, Greece

¹²Harvard University, Cambridge, MA 02138, USA

¹³University of New Mexico, Albuquerque, NM 87131, USA

¹⁴Makerere University and Infectious Diseases Research Collaboration (IDRC), Kampala, Uganda

VEuPathDB Website Privacy Policy

UPDATED: April 12, 2020

Introduction

VEuPathDB (also referred to as "we" throughout this document) is committed to protecting its users' 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 community'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 "referrer" 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 "Contact Us" 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.

"Contact Us" Form

The header on each web page includes a "Contact Us" 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 browser'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 user's privacy settings, this information may be displayed on the user's profile page or in the user's tweets which may be retweeted on VEuPathDB'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 Twitter'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 Twitter'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 user's privacy settings, this information may be displayed on the user's profile page along with any activity such as comments or "likes" 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 [Google'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 "Analyze My Experiment" 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 [Globus's 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 specific 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 tutorials

Site Search

(updated: 15 Apr 2020)

Explore site search and learn how to filter your results or export them to a search strategy

- [SiteSearch.pdf](#)

Genome Browser

(updated: 15 Apr 2020)

Basic tutorial on how to effectively explore data in the genome browser including track selection and configuration.

- [JBrowseBasics.pdf](#)

Regular expressions & colocation

(updated: 15 Apr 2020)

Use regular expressions to identify protein or DNA patters. Use the colocation tool to identify DNA segments based on their relative location on the genome.

- [RegEXcolocate.pdf](#)

Variant Calling

(updated: 15 May 2020)

Analyze DNA sequence for variants using a workflow in VEuPathDB Galaxy

- [VariantAnalysisVEuPathDBGalaxyPart1.pdf](#)

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
<div><div></div><div>(Book Chapter) Omar Harb, Jessica C. Kissinger and David S. Roos on behalf of the EuPathDB group ToxoDB: the functional genomic resource for <i>Toxoplasma</i> <i>Toxoplasma gondii</i> 3rd edition Edited by Louis Weiss and Kami Kim (2020) https://doi.org/10.1016/B978-0-12-815041-2.00023-2</div></div>
<div><div></div><div>(Book Chapter) Susanne Warrenfeltz and Jessica Kissinger on behalf of the EuPathDB Consortium Accessing <i>Cryptosporidium</i> 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</div></div>
<div><div></div><div>(Book Chapter) Omar S. Harb and David S. Roos on behalf of the EuPathDB Consortium ToxoDB: Functional Genomics Resource for <i>Toxoplasma</i> and Related Organisms Methods in Molecular Biology Editor, Christopher J. Tonkin (2020) https://doi.org/10.1007/978-1-4939-9857-9_2</div></div>

(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-CalderÃ³n 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-Calder n 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, Niernan 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), D729  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), 42  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.    2011 by John Wiley & Sons, Inc. PMID: 21901743
[[Abstract](#) | [Full text](#) | [PDF](#)]

Aurrecochea 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, Aurrecochea 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](#)

Aurrecochea 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, Aurrecochea 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), 308-313

<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., & Collins, F. H.

VectorBase: a data resource for invertebrate vector genomics

Nucleic acids research, 1 January 2009; 37(Database issue), D583-D587

<https://doi.org/10.1093/nar/gkn857>

Aurrecochea, 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>

Aurrecochea, 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., & 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 Aurrecochea, 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 Aurrecochea, 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, Aurrecochea 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), 893-899
<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 Webinars

Upcoming Webinars

Thursday July 23th, 2020: [Variant data in VEuPathDB.](#)

Previous Webinars

If you missed any of our previous webinars feel free to explore the free recordings:

Thursday July 16th, 2020: [Introduction to FungiDB - Data and Functionality.](#)

Thursday July 9th, 2020: [RNAseq searches in VEuPathDB.](#)

Thursday July 2nd, 2020: [Orthology searches and OrthoMCL.](#)

Thursday June 25th, 2020: [Searching Phenotypic datasets in VEuPathDB.](#)

Thursday June 18th, 2020: [Special event: Introduction for VectorBase users.](#)

Thursday June 11th, 2020: [MapVEu: population biology map filtering and searches.](#)

Wednesday June 10th, 2020: [VEuPathDB Release 47.](#)

Thursday June 4th, 2020: [Motif searches and regular expressions in VEuPathDB.](#)

Thursday May 28th, 2020: [GO and other enrichment analyses in VEuPathDB.](#)

Thursday May 21st, 2020: [Manual gene annotation in Apollo.](#)

Thursday May 14th, 2020: [Running a Galaxy workflow and integrating data into VEuPathDB.](#)

Thursday May 7th, 2020: [Building search strategies in VEuPathDB.](#)

Thursday April 30th, 2020: [Interpreting RNAseq data in the genome browser.](#)

Thursday April 23rd, 2020: [Introduction to the genome browser in VEuPathDB.](#)

Thursday April 16th, 2020: [Introduction to VEuPathDB resources.](#)

Upcoming Webinars

Topic: Variant data in VEuPathDB
Date: Thursday July 23th, 2020
Time: 10AM US Eastern Time
Registration Link: <https://attendee.gotowebinar.com/register/9005464633523261196>
Additional Details:
This webinar will cover:

- Single nucleotide polymorphism searches
- Copy number variation searches
- Variant data visualizations

Previous Webinars

Topic: Introduction to FungiDB - Data and Functionality
Date: Thursday July 16th, 2020
Time: 10AM US Eastern Time
Recording Link: https://youtu.be/EAL_7uCPH3Q
Handout: [Anatomy of the main page](#)
Additional Details:
This webinar will cover:

- What type of data is in FungiDB
- How do we find out about data and schedule it for release
- Demo of key searches
- Demo of key search and analysis tools

Topic: RNAseq searches in VEuPathDB
Date: Thursday July 9th, 2020
Time: 10AM US Eastern Time
Recording Link: <https://youtu.be/D2NAqIz6Fug>
Additional Details:
This webinar will demonstrate how to search for under differential expression with RNAseq datasets.

Topic: Orthology searches and OrthoMCL
Date: Thursday July 2nd, 2020
Time: 10AM US Eastern Time
Recording Link: <https://youtu.be/KYD44dBteSs>
Additional Details:
This webinar will demonstrate how to search for homolog genes.

Topic: Searching Phenotypic datasets in VEuPathDB
Date: Thursday June 25th, 2020
Time: 10AM US Eastern Time

Recording Link: <https://youtu.be/0KNatBANi94>

Additional Details:

This webinar gives an overview of phenotypic data on VEuPathDB sites, including representation of data on gene pages and searches that access all data so that you can find all genes with a certain phenotype. Phenotype data on our sites is based on cellular imaging experiments, and mutagenesis studies (among others).

Topic: Special event: Introduction for VectorBase users

Date: Thursday June 18th, 2020

Time: 10AM US Eastern Time

Recording Link: <https://youtu.be/nETsJDwFxBA>

Additional Details:

Topics covered included:

- Tour of the new (beta) website layout and organization
 - Overview of the new tools
 - Process for new data set incorporation
-

Topic: MapVEu: population biology map filtering and searches

Date: Thursday June 11th, 2020

Time: 10AM US Eastern Time

Recording link: <https://youtu.be/XhnUuDN7CLO>

Additional Details:

This webinar will demonstrate how to use the MapVEu tool in VEuPathDB. Topics covered will include:

- Basic MapVEu navigation
 - Filtering data in the map
 - Exploring record details
-

Topic: VEuPathDB Release 47

Date: Wednesday June 10th, 2020

Time: 10AM US Eastern Time

Recording link: <https://youtu.be/Ab0V1ojYono>

Additional Details:

This webinar will discuss and demonstrate new features and data available in release 47 of the newly merged VEuPathDB sites.

Topic: Motif searches and regular expressions in VEuPathDB

Date: Thursday June 4th, 2020

Time: 10AM US Eastern Time

Recording link: <https://youtu.be/kaQbEdtjIew>

Handout: [Motifs and regular expressions 1](#)

Handout: [Motifs and regular expressions 2](#)

Handout: [Motifs and regular expressions 3](#)

Handout: [Motifs and regular expressions presentation slides](#)

Additional Details:

This webinar will describe the different available motif searches including regular expressions in VEuPathDB. Topics covered will include:

- InterPro domains
 - InterPro domain searches
 - Regular expressions
 - Regular expressions searches
-

Topic: GO and other enrichment analyses in VEuPathDB

Date: Thursday May 28th, 2020

Time: 10AM US Eastern Time

Recording Link: <https://youtu.be/15Uf49T9AF4>

Handout: [Enrichment Analysis in VEuPathDB](#)

Additional Details:

This webinar will guide you through the process of running enrichment analyses in VEuPathDB. Topics covered will include:

- Gene ontology
 - GO enrichment
 - Metabolic pathway enrichment
-

Topic: Manual gene annotation in Apollo

Date: Thursday May 21st, 2020

Time: 10AM US Eastern Time

Recording Link: <https://youtu.be/RMXZDBMh8Fo>

Additional Details:

This webinar will guide you through the process of updating and fixing gene models based on underlying data. Topics covered will include:

- How to add new gene models
 - How to alter the structure of a gene
 - How to use underlying data to support structural changes
-

Topic: Running a Galaxy workflow and integrating data into VEuPathDB

Date: Thursday May 14th, 2020

Time: 10AM US Eastern Time

Recording Link: <https://youtu.be/k15AJZpNH8>

Handout: [How to set up an account in VEuPathDB Galaxy](#)

Handout: [RNA sequence data analysis in VEuPathDB Galaxy Part 1](#)

Handout: [RNA sequence data analysis in VEuPathDB Galaxy Part 2](#)

Additional Details:

This webinar will guide you through the process of running a workflow in galaxy and exporting the results to your VEuPathDB workspace.

Topic: Building search strategies in VEuPathDB

Date: Thursday May 7th, 2020

Time: 10AM US Eastern Time

Recording Link: <https://youtu.be/hmTzUUibSeY>

Handout: [Building Strategies- example in PlasmoDB](#)

Handout: [Building search strategies in FungiDB](#)

Additional Details:

This webinar will guide you through the process of building search strategies in VEuPathDB resources. Topics covered will include:

- Running searches
 - Combining searches in your strategy space
 - Editing steps in your search strategy
 - Examining results of search strategies
 - Analyzing search strategy results (eg. GO enrichment)
-

Topic: Interpreting RNAseq data in the genome browser

Date: Thursday April 30th, 2020

Time: 10AM US Eastern Time

Recording Link: https://youtu.be/W-HKOMq2_tl

Additional Details:

This webinar will guide you through the process of viewing RNA-seq coverage tracks in the genome browser and how to use this data to examine gene structure. Specifically the webinar will cover:

- Loading RNA-seq and viewing data tracks
 - Loading and viewing splice junction tracks
 - Interpreting RNA-seq results to confirm gene models and propose alternative ones
-

Topic: Introduction to the genome browser in VEuPathDB

Date: Thursday April 23rd, 2020

Time: 10AM US Eastern Time

Recording Link: <https://youtu.be/GfV1LrycO7Y>

Additional Details:

This webinar will introduce you to the genome browser in VEuPathDB resources. Topics covered will include:

- How to get started with JBrowse
 - How to add tracks
 - How to search in JBrowse
 - How to move around
 - How to load your own data
-

Topic: Introduction to VEuPathDB resources

Date: Thursday April 16th, 2020

Time: 10AM US Eastern Time

Recording Link: <https://youtu.be/YprhSDTsEE>

Additional Details:

This webinar will introduce you to VEuPathDB resources and highlight some of the tools available to you to explore and analyze underlying data. Topics covered will include:

- How to get started
- RNAseq searches
- MapVEu
- Phenotype data
- Galaxy tools

VEuPathDB Workshops

Upcoming Workshops

2020 VectorBase, Arthropod Genomics Symposium (AGS), virtual.

- [VectorBase workshop at AGS](#), on Friday July 24th, 2020 (3-hour event)
- [Gene manual annotation with Apollo workshop at AGS](#), on Monday July 27th, 2020 (2-hour event)

Previous Workshops

VectorBase: [Previous tutorials](#) and [Previous workshops and webinars](#).

EuPathDB: [Previous workshops](#).

Upcoming Workshops

VectorBase workshop at AGS , on Friday July 24th, 2020

The VectorBase workshop is a 3-hour introduction to this database, which is part of the VEuPathDB project.

Who should apply: The ideal candidate has a need to use VectorBase (or any of the other parasite and fungi databases, which all have the same layout) and presumably has a bioinformatics research problem to work on after attending the course.

Database covered: VectorBase

Costs and Fees: The course is free. However, to make the most out of it we will have limited seats in order to ensure participants can easily interact with the instructors while solving the exercises. Please only apply if you plan to attend the workshop in full

Where: Accepted participants will receive the appropriate workshop link

Application deadline: Friday, July 17th 2020

Acceptance notification by: Monday, July 20th 2020

Registration link: https://docs.google.com/forms/d/e/1FAIpQLSeEpeQkQwnV59EG_KSH93PeI7Zb6f_5L-4po6FZ_jo-EKX8bA/viewform

Note: To encourage the active participation/Q&A during the hands-on activities, limited seats are available and the sessions will not be recorded. Priority seating will be given to AGS meeting attendees.

▼ Agenda, schedule and instructors:

- GIGC: Gloria I. Giraldo-Calder³ⁿ, University of Notre Dame
- OH: Omar Harb, Pennsylvania State University
- Other members from the VEuPathDB team will join us to help you during the exercises

Time				Topic	Instructor
PDT (W. Coast)	EDT (E. Coast)	CET (Europe)	Length (min)		
8:00	11:00	17:00	10	Quick introduction and demonstrations about the VEuPathDB resources	GIGC
8:10	11:10	17:10	20	1. Site search: Exercise and Answer key	GIGC
8:30	11:30	17:30	40	2. Search strategies: Exercises and Answer key	GIGC

Time				Topic	Instructor
PDT (W. Coast)	EDT (E. Coast)	CET (Europe)	Length (min)		
			15	Break	
9:25	12:25	18:25	40	3. GO enrichment: Exercises and Answer key	OH
10:10	13:10	19:10	40	4. Gene page and genome browser: Exercises and Answer key	GIGC
10:50	13:50	19:50	15	Additional questions, wrap-up and workshop feedback	OH

Additional Information: Email your questions to help@vectorbase.org

Gene manual annotation with Apollo workshop at AGS , on Monday July 27th, 2020

This workshop is a 2-hour introduction to the tool Apollo, which is currently available in VectorBase. This course will guide you through the process of updating and fixing gene models based on underlying data, including topics such as the addition of new genes, edition of the intron/exon boundaries, use of RNAseq and homology alignment data to support changes. This workshop would benefit vector biologist or any scientist with interest in annotating their favorite genes in Apollo available in other databases, e.g., i5k (<https://i5k.nal.usda.gov/jbrowse-and-apollo>).

Who should apply: The ideal candidate has a need to use Apollo and presumably has a list of genes to work on after attending the course.

Tool covered: VectorBase Apollo

Costs and Fees: The course is free. However, to make the most out of it we will have limited seats in order to ensure participants can easily interact with the instructors while solving the exercises. Please only apply if you plan to attend the workshop in full

Where: Accepted participants will receive the appropriate workshop link

Application deadline: Friday, July 17th 2020

Acceptance notification by: Monday, July 20th 2020

Registration link: https://docs.google.com/forms/d/e/1FAIpQLSeEpeQkQwnV59EG_KSH93PeI7Zb6f_5L-4po6FZ_jo-EKX8bA/viewform

Note: To encourage the active participation/Q&A during the hands-on activities, limited seats are available and the sessions will not be recorded. Priority seating will be given to AGS meeting attendees.

▼ Agenda, schedule and instructors:

- CH: Christopher Childers, United States Department of Agriculture
- GIGC: Gloria I. Giraldo-Calder  n, University of Notre Dame
- DL: Daniel Lawson, Imperial College London
- OH: Omar Harb, Pennsylvania State University
- Other members from the VEuPathDB team will join us to help you during the exercises

Time				Topic	Instructor
PDT (W. Coast)	EDT (E. Coast)	CET (Europe)	Length (min)		
8:00	11:00	17:00	15	i5K@NAL: Who we are and what data we host with an emphasis on Apollo	CH
8:15	11:15	17:15	45	Demonstration	DL
			15	Break	
9:15	12:15	18:15	45	Hands-on exercises	DL, OH, GIGC

Additional Information: Email your questions to help@vectorbase.org

Previous Workshops

VectorBase: [Previous tutorials](#) and [Previous workshops and webinars](#).

