Nucleic Acids Research, 2026, gkag1245
https://doi.org/10.1093/nar/gkag1245
Database issue

VEuPathDB: 2026 update on data, tools, and infrastructure for eukaryotic pathogens, vectors, and hosts

Cristina Aurrecoechea[1], Evelina Basenko[2,3], Daniel P. Beiting[4], Ulrike Böhme[2,3], John Brestelli[5], Stuart Brown[5], Kathryn Crouch[6], Richard Demko[5], Ryan Doherty[5], Luc Elliott[2,3], Steve Fischer[5], Bindu Gajria[5], Omar S. Harb[5,*], Elizabeth Harper[5], Danica Helb[5], Sufen Hu[5], Andrew R. Jones[2,3], Jessica C. Kissinger[1], Nupur Kittur[1], Wei Li[5], Adriana Ludwig[2,3], Robert M. MacCallum[7], Antony McCabe[3,8], Mary Ann McDowell[9], Shahram Mesdaghi[2,3], Ahmed Muharram[5], Sadik Muzemil[2,3], Mustafa Veysi Nural[1], Freyja Redmond[2,3], Daniel J. Rigden[2,3], David S. Roos[5], Drew Spruill[1], Jordan Tzvetkov[2,3], Mariann Winkelman[5]

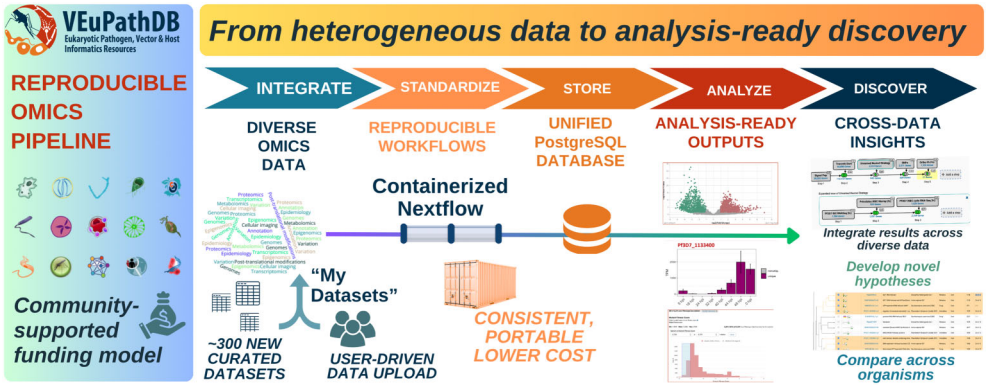
[1]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA, 30602, United States
[2]Department of Biochemistry, Cell and Systems Biology, University of Liverpool, Liverpool, England, L69 7BE, United Kingdom
[3]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, England, L69 7BE, United Kingdom
[4]School of Veterinary Medicine, Department of Pathobiology, University of Pennsylvania, Philadelphia, PA, 19104-4539, United States
[5]Department of Biology, University of Pennsylvania, Philadelphia, PA, 19104-6017, United States
[6]Wellcome Centre for Integrative Parasitology, University of Glasgow, Glasgow, Scotland, G12 8TA, United Kingdom
[7]Imperial College London, London, England, SW7 2AZ, United Kingdom
[8]Computational Biology Facility, University of Liverpool, Liverpool, England, L69 7ZB, United Kingdom
[9]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, 46556, United States

*To whom correspondence should be addressed. Email: oharb@sas.upenn.edu

Abstract

The eukaryotic pathogen, vector, and host informatics resource (VEuPathDB, https://veupathdb.org) provides open access to genomic and other large-scale data for protozoan and fungal pathogens, arthropod vectors, and their hosts, supporting >800 organisms and 2602 datasets across 12 component databases, OrthoMCL, and the newly added dataExplorer, as of Release 71 (August 2026). Since our last update, 292 datasets have been integrated, gene product descriptions improved, and OrthoMCL expanded to over 819 000 orthology groups spanning 9.2 million proteins. Artificial intelligence-assisted approaches now support dataset curation, community gene annotation, and experimental summaries of transcriptomic evidence on gene pages, each subject to expert or user review. The new My Datasets workspace lets researchers upload, analyze, and share their own data alongside curated datasets, and a guided interface supports differential expression analysis without programming. Infrastructure has been consolidated from 15 Oracle databases into a single PostgreSQL platform, with data-processing pipelines migrated to containerized Nextflow workflows. This period was also defined by the loss of the long-standing US federal award that funded VEuPathDB, its temporary closure in September 2024, and its restoration under a community-supported model of voluntary tiered subscriptions, donations, and philanthropic support, with all data and tools remaining freely accessible.

Graphical abstract



Received: September 15, 2026. **Revised:** October 26, 2026. **Accepted:** October 27, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

Introduction

The eukaryotic pathogen, vector and host informatics resource (VEuPathDB, https://veupathdb.org) provides free, open access to genomic and other large-scale datasets for eukaryotic protozoan and fungal pathogens, arthropod vectors of disease, and their hosts. VEuPathDB comprises 12 organism-specific component databases together with the OrthoMCL ortholog resource and the newly added dataExplorer, all accessible through the umbrella VEuPathDB portal (Table 1), and collectively supports >800 organisms and 2602 integrated datasets as of Release 71 in August 2026. Organisms in VEuPathDB encompass a wide array of important pathogens and vectors, affecting humans, livestock and companion animals, and major crops. The diseases involved include malaria, toxoplasmosis, cryptosporidiosis, giardiasis, trichomoniasis, amoebiasis, microsporidiosis, leishmaniasis, Chagas disease and sleeping sickness, babesiosis and theileriosis in livestock, and invasive and opportunistic fungal infections. VEuPathDB also includes species on the World Health Organization (WHO) fungal priority pathogens list, including the critical-group pathogens *Aspergillus fumigatus*, *Candida*, and *Cryptococcus*, and agents of WHO-designated neglected tropical diseases. (For the full listing of organisms and diseases studied across VEuPathDB, see https://veupathdb.org/veupathdb/app/static-content/disease-research.html). Their impact extends beyond clinical disease to crop losses and food security, livestock productivity, and biosecurity preparedness—a One Health remit that makes shared, cross-taxon data infrastructure particularly valuable. Research on these organisms generates heterogeneous omics data that are dispersed across repositories and generated under differing conditions, and that few individual laboratories have the resources to integrate and analyze at scale. VEuPathDB exists to close that gap, bringing these data together under consistent, standardized workflows and making them findable, interoperable, and reusable through a common set of web-based search, visualization, and analysis tools.

VEuPathDB integrates a broad range of data types—genome sequence and annotation, transcriptomics, proteomics, epigenomics, metabolomics, phenotype and functional-genomic screens, population genetic variation, and clinical and host-response data—all processed through standardized workflows, including orthology prediction with OrthoMCL, so that results are directly comparable across datasets, data types, and organisms [1, 2]. Rather than hosting these data as static downloads, VEuPathDB presents them through a graphical search system in which individual searches are combined into multi-step search strategies, allowing sets of genes, isolates, or samples to be built, intersected, transformed, and refined without programming. Results link to record pages for individual genes and other entities and to integrated tools, including sequence similarity search, the genome browser, and community annotation through Apollo. All of this runs in the browser, with no local installation or programming required. These established capabilities are described in detail in previous database updates and in step-by-step protocols for their use [3], and affiliated resources built on the same infrastructure extend the model to other data domains, including clinical and epidemiological studies in ClinEpiDB [4] and microbiome studies in MicrobiomeDB [5]. VEuPathDB serves ∼46 000 unique users per month across >170 countries, and has close to 40 000 Google Scholar citations (https://veupathdb.org/veupathdb/app/static-content/pubsUsingVEuPath.html). It is designated a Global Core Biodata Resource by the Global Biodata Coalition (https://globalbiodata.org/what-we-do/global-core-biodata-resources/list-of-current-global-core-biodata-resources/) and an ELIXIR Core Data Resource (https://elixir-europe.org/platforms/data/core-data-resources). The present update therefore concentrates on what has changed since Release 66 in November 2023 and up to release 71 in August 2026.

The period covered by this update was shaped by a fundamental change in how VEuPathDB is funded. In 2024, after two decades of continuous support, the US National Institute of Allergy and Infectious Diseases did not renew the award that had underwritten the resource, and VEuPathDB was taken offline on 15 September 2024 [6]. The prospect of losing it prompted a broad international response: 43 members of the National Academies of Sciences, Engineering and Medicine wrote to the Director of the National Institutes of Health asking that the resource be kept live until a proven alternative existed, >1000 researchers wrote to the funding agency, and news coverage and commentary in Science, Nature, the BugBitten blog, and The Lancet Microbe set out what the closure of VEuPathDB would mean for parasitology, mycology, and vector biology—particularly in disease-endemic countries, where VEuPathDB underpins both primary research and local training capacity [6–9]. Philanthropic bridge funding allowed the resource to be restored, initially in static form, while a new community-supported funding model was established (see section below—New Business Model). Bimonthly releases were paused after Release 68 in September 2024 and resumed with Release 69 in March 2026. During this period the infrastructure, data-processing pipelines, and curation workflows described below were re-engineered for a smaller, more sustainable operating footprint, while preserving the depth of integration and expert curation on which users depend.

Since Release 66, 292 new datasets have been integrated, alongside improved gene product descriptions and a major update to OrthoMCL. The underlying infrastructure has been modernized: siloed Oracle databases have been consolidated into a single PostgreSQL platform, and genome and proteome data-processing workflows have been migrated to publicly available, containerized Nextflow pipelines, improving reproducibility and portability while reducing cost. Artificial intelligence (AI)-assisted approaches have been introduced to facilitate dataset curation, with careful review of AI-drafted content by a curator before integration and release. AI summarization of transcriptomic evidence on gene pages, which is generated automatically and presented to users as experimental. The user-comment system, through which researchers have contributed thousands of gene-level comments for more than two decades, has a redesigned interface that adds an AI-assisted summaries of user provided documents, with drafts reviewed and approved by the commenter before publication. Separately, the "My Datasets" workspace allows data providers to upload, manage, and share (if they wish) their own datasets, complementing staff-led integration. Release-by-release detail on all of these developments is available on the VEuPathDB news page (https://veupathdb.org/veupathdb/app/static-content/VEuPathDB/news.html).

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

Table 1. Overview of VEuPathDB component databases and resources

Project	Web address	URL to access list of organisms supported	Number of datasets (release 71)	General description of organisms and disease
VEuPathDB	https://veupathdb.org	https://veupathdb.org/veupathdb/app/search/organism/GenomeDataTypes/result	2602	Umbrella resource for all component databases; supports eukaryotic protozoan and fungal pathogens, disease vectors, and their hosts
AmoebaDB	https://amoebadb.org	https://amoebadb.org/amoeba/app/search/organism/GenomeDataTypes/result	88	Supports *Entamoeba* and other amoebae, including model organisms and species that cause amoebiasis
CryptoDB	https://cryptodb.org	https://cryptodb.org/cryptodb/app/search/organism/GenomeDataTypes/result	88	Supports *Cryptosporidium* species that cause cryptosporidiosis
FungiDB	https://fungidb.org	https://fungidb.org/fungidb/app/search/organism/GenomeDataTypes/result	764	Supports fungi and fungus-like oomycetes, including species that cause invasive and opportunistic fungal infections
GiardiaDB	https://giardiadb.org	https://giardiadb.org/giardiadb/app/search/organism/GenomeDataTypes/result	54	Supports *Giardia* species that cause giardiasis
HostDB	https://hostdb.org	https://hostdb.org/hostdb/app/search/organism/GenomeDataTypes/result	86	Supports host organisms (e.g. human and other vertebrate genomes) used as reference genomes for pathogen and vector host studies
MicrosporidiaDB	https://microsporidiadb.org	https://microsporidiadb.org/micro/app/search/organism/GenomeDataTypes/result	65	Supports microsporidia, opportunistic pathogens that cause microsporidiosis
PiroplasmaDB	https://piroplasmadb.org	https://piroplasmadb.org/piro/app/search/organism/GenomeDataTypes/result	49	Supports piroplasms such as *Babesia* and *Theileria*, which cause babesiosis and theileriosis, mainly in livestock
PlasmoDB	https://plasmodb.org	https://plasmodb.org/plasmo/app/search/organism/GenomeDataTypes/result	325	Supports organisms that cause malaria
ToxoDB	https://toxodb.org	https://toxodb.org/toxo/app/search/organism/GenomeDataTypes/result	188	Supports *Toxoplasma* and related coccidian parasites that cause toxoplasmosis and other diseases
TrichDB	https://trichdb.org	https://trichdb.org/trichdb/app/search/organism/GenomeDataTypes/result	26	Supports *Trichomonas* species that cause trichomoniasis
TriTrypDB	https://tritrypdb.org	https://tritrypdb.org/tritrypdb/app/search/organism/GenomeDataTypes/result	281	Supports kinetoplastid parasites such as *Trypanosoma* and *Leishmania*, which cause sleeping sickness, Chagas disease, and leishmaniasis
VectorBase	https://vectorbase.org	https://vectorbase.org/vectorbase/app/search/organism/GenomeDataTypes/result	615	Supports arthropod vectors, such as mosquitoes, ticks, and other disease-transmitting insects
OrthoMCL DB	https://orthomcl.org	https://orthomcl.org/orthomcl/app/release-summary	857 genomes represented (a)	Provides cross-genome ortholog groupings for eukaryotic and prokaryotic organisms; not organism- or disease-specific
dataExplorer	https://dataExplorer.org/	https://dataExplorer.org/dexp/app/search/dataset/AllDatasets/result	9 public user datasets	Tool designed for quick data exploration and visualization of data tables.

(a) Includes core and peripheral organisms.
This table lists VEuPathDB and its component databases, each dedicated to a specific group of eukaryotic pathogens, disease vectors, or host organisms. For more information on the diseases studied across VEuPathDB, see https://veupathdb.org/veupathdb/app/static-content/disease-research.html.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

New in VEuPathDB

New business model

Before adopting a new funding model, VEuPathDB consulted its user community directly. From 14 September 2024 to 7 January 2025, VEuPathDB conducted a User Impact and Sustainability Survey. The survey collected information on reliance on VEuPathDB resources, patterns of database use, the features users consider essential, and attitudes toward alternative funding mechanisms, and it received 1862 responses. A full report, including an executive summary and analysis of both quantitative and qualitative responses, and a shorter summary are publicly available (https://static-content.veupathdb.org/documents/PUBLIC_REPORT_VEuPathDB_User_Impact_Sustainability_Survey.pdf;https://static-content.veupathdb.org/documents/Survey_Summary.pdf).

Respondents expressed widespread concern about the prospect of losing the resource, which they described as effectively irreplaceable for research on fungi, protozoan parasites, and arthropod disease vectors, and on the human, animal, and agricultural diseases they cause—many of them globally important yet neglected. While users were broadly open to alternative funding mechanisms, the community expressed a strong preference for preserving broad accessibility, indicating that any model adopted would need to avoid excluding researchers on the basis of ability to pay.

These responses also reflect the role VEuPathDB has played over many years. The resource has grown from a small number of databases for apicomplexan parasites to encompass a broad range of protozoan parasites, fungi and oomycetes, arthropod and mollusc vectors, and their hosts, and has provided a long-term home for research communities whose dedicated resources would have been difficult to sustain independently. By bringing these communities together under one uniform umbrella, with shared infrastructure, consistent data processing and curation standards, and a common set of search and analysis tools, the VEuPathDB model has made it possible to support multiple underserved research communities efficiently, while allowing researchers to compare data across organisms within a single resource.

The model that emerged reflects that constraint. Beginning in March 2025, VEuPathDB introduced a voluntary subscription service intended to sustain operations while keeping the resource open and accessible (https://veupathdb.org/veupathdb/app/static-content/subscriptions.html). Registration with the site is required of users, but the decision to subscribe is left to them: access is not gated on payment, and the model relies on the good faith of the user community. This arrangement is consistent with current "Nucleic Acids Research" (NAR) policy: while the journal has historically required that the content of contributing resources be freely available without user registration, the 2026 Database Issue editorial introduces flexibility in exceptional cases, noting that databases forced by acute funding problems to charge users and to introduce necessary user registration may still be invited to submit by prior agreement with the editor [10]. Registration serves two further purposes. It allows usage to be attributed to identifiable research groups and institutions, which makes group-level subscription possible, and yields more accurate statistics on who relies on the resource and how they use it—information that is essential both to setting fee levels appropriately and to identifying where sustainability efforts should be directed. Registered users can save searches, data mining results, and preferences, and have access to a growing set of custom tools for upload and analysis of their own data under My Datasets (see below). Subscription fees are tiered rather than uniform: they were established through consultation with the user communities and are scaled to the size of a group's operation and the intensity of its use for data mining and dissemination, with a subsidized tier at no cost for groups that cannot pay. Subscribing groups are acknowledged publicly (with permission) on the VEuPathDB site, and voluntary donations provide an additional route of support.

As of 1 September 2026, VEuPathDB had 38 121 registered users, of whom 3397, drawn from 567 research groups, had subscribed. In the first nine months of operation, subscriptions supplied ∼25% of the resource's operating budget.

The survey's central conclusion—that a sustainable future is likely to require a combination of equitable fee structures, diversified funding sources, and reduced operating costs—maps onto the strategy that VEuPathDB has followed. The subscription service and donations address the first two elements; the infrastructure and workflow changes described below address the third, lowering the recurring cost of operating the resource so that a community-supported revenue base is sufficient to maintain it.

New datasets

VEuPathDB continues to expand the breadth and depth of data available to researchers studying eukaryotic pathogens, their hosts, and vectors. Recent releases have added new genomes, bulk and single-cell RNAseq datasets, proteomic and epigenomic studies, population and phenotype data, and host pathogen interaction data. Since Release 66 (November 2023) through Release 71 (August 2026), 292 new datasets have been integrated across the resource, with release-by-release detail available on the VEuPathDB news page. A few examples given here highlight the range of these additions. The dataset pages on the VEuPathDB site (https://veupathdb.org/veupathdb/app/search/dataset/AllDatasets/result) provide full citations and provenance for every dataset loaded.

Expanding support for phytopathogens

A newly funded initiative is extending VEuPathDB's coverage of fungal and oomycete plant pathogens, in close collaboration with Ensembl. FungiDB already integrates data for many economically important phytopathogens—including species of *Blumeria*, *Botrytis*, *Colletotrichum*, *Fusarium*, *Pyricularia*, *Melampsora*, *Phytophthora*, *Puccinia*, *Ustilago*, and *Zymoseptoria*—but until now this support has not been a funded priority. A dedicated project funded by the Biotechnology and Biological Sciences Research Council is addressing that gap: loading additional phytopathogen genomes and accompanying functional genomic data, generating cross-taxon orthology mappings and protein domain annotations, and deploying fungal-specific tools targeting virulence factors and biosynthetic gene clusters. HostDB is being expanded in parallel to include plant host genomes, allowing pathogen and host data to be queried together—for example, identifying rice genes with altered expression during *Pyricularia oryzae* infection—using the Search Strategy system and dual-transcriptome tools already built into VEuPathDB.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

Genomes

New genomes and annotation updates span all major taxonomic groups represented in the resource. Eleven new apicomplexan genomes were added: four *Cryptosporidium* genomes in CryptoDB, five *Eimeria* genomes in ToxoDB, and one genome each for *Plasmodium* in PlasmoDB and *Babesia* in PiroplasmaDB. TriTrypDB added two new kinetoplastid assemblies, one of them a hybrid long- and short-read assembly, together with an updated structural annotation for the *Trypanosoma brucei* reference genome. AmoebaDB updated one amoebozoan reference annotation and released a second as a pre-publication genome-browser track. FungiDB added seven new Ascomycota genomes spanning the Dothideomycetes, Eurotiomycetes, and Leotiomycetes, along with three new oomycete assemblies, and replaced 14 existing reference genomes with improved assemblies, comprising nine Ascomycota and five Basidiomycota, the latter predominantly rust fungi; one further structural annotation was updated. VectorBase added 20 vector and intermediate-host assemblies: nine mosquito genomes, several from genera not previously represented in the resource; seven other dipteran genomes covering tsetse, stable, house, blow, and sand flies; two tick genomes; one flea genome; and one freshwater snail genome from a genus that serves as an intermediate host in schistosomiasis transmission. HostDB added one new mammalian host genome.

Bulk and single-cell transcriptomics data

Transcriptomics accounts for the largest share of new data, with roughly 150 bulk RNA-seq experiments added across the resource. Apicomplexan datasets predominate among the parasite additions, covering life-cycle progression and sexual development, stage-specific and gametocyte transcriptomes, liver-stage and cytoadherent parasite populations, and small-RNA profiling, distributed across PlasmoDB, ToxoDB, CryptoDB and PiroplasmaDB. TriTrypDB added kinetoplastid transcriptomes comparing sequestered and non-sequestered parasite populations, and AmoebaDB added developmental-program transcriptomes for two amoebozoan species. FungiDB integrated the largest number of studies, spanning human fungal pathogens, dermatophytes, an amphibian pathogen and a wide range of plant-pathogenic fungi and oomycetes, with designs covering antifungal drug response, spore germination, host colonization and transcription-factor mutants. VectorBase added arthropod and mollusc datasets addressing life stages, blood feeding, temperature response, ivermectin exposure and region-specific gut transcriptomes across mosquitoes, tsetse and sand flies, triatomines, ticks, lice, mites, and snails. HostDB contributed host-side transcriptomes of responses to fungal and parasitic infection in human peripheral blood mononuclear cells, macrophage and dendritic-cell populations, fibroblasts, and patient skin biopsies.

Single-cell and spatial approaches expanded substantially over this period. ToxoDB now hosts a paired single-cell RNA-seq and assay for transposase-accessible chromatin sequencing (ATAC-seq) atlas of the *Toxoplasma* lytic cycle, providing complementary views of transcription and chromatin accessibility, while paired PlasmoDB and VectorBase datasets map *Plasmodium* transitions and interactions within the female *Anopheles* from both parasite and vector perspectives, complemented by single-cell and spatial transcriptomics of ∼50 000 *Anopheles* brain cells.

Because single-cell and spatial datasets are expected to grow rapidly, VEuPathDB is developing tools to support their exploration and integration with other data in the resource. Planned features include visualization of spatial transcriptomics data alongside single-cell profiles, and interactive selection of user-defined cell populations, for example on a uniform manifold approximation and projection plot, for differential expression analysis, with the resulting gene lists available for use in search strategies. We are also exploring AI-assisted approaches to help users interpret cell populations and the genes that distinguish them. Support for single-cell RNA-seq data in My Datasets is also planned, so that researchers can analyze their own single-cell results alongside existing data. A key obstacle to integration is that commonly used single-cell analysis tools often modify gene identifiers during processing (for example, by replacing underscores with hyphens), so that results can no longer be mapped back to genes in VEuPathDB; we are developing approaches to preserve or restore these identifiers so that single-cell data can be integrated more reliably.

Proteomics and spatial biology

Proteomic additions extend across parasite, fungal and host systems. PlasmoDB added the translatome of *Plasmodium falciparum* mature stage V gametocytes, a *P. falciparum* phosphorylation meta-analysis, and the whole-cell spatial hyperLOPIT proteome of the *P. falciparum* 3D7 schizont, which resolves protein localization at subcellular scale. ToxoDB added a *Toxoplasma gondii* phosphorylation meta-analysis and comparative proteomics of SAG1 knockout clones, and TriTrypDB added a *Trypanosoma brucei* phosphoproteome meta-analysis. FungiDB integrated *Aspergillus* conidial surface proteomes, and HostDB added the membrane proteome of *Cryptosporidium parvum*-infected HCT-8 cells.

Epigenomics and chromatin biology

New chromatin-focused datasets broaden the resource beyond gene expression to the regulation of the genome. TriTrypDB added a chromatin immunoprecipitation sequencing (ChIP-seq) survey of 69 chromatin-associated factors, constituting a comprehensive epigenomic resource for kinetoplastids. FungiDB added H3K9 acetylation ChIP-seq together with paired ATAC-seq and ChIP-seq of *Phytophthora infestans* T30-4 spanning mycelium and two infection stages. The ToxoDB single-cell ATAC-seq map of the *Toxoplasma gondii* lytic cycle, noted above, is the first single-cell chromatin dataset in the resource.

Genetic variation and functional genomics

New resources also capture genetic diversity and experimentally determined gene function. FungiDB added whole-genome single-nucleotide polymorphism (SNP) datasets for seven fungal species, spanning human and crop pathogens, together with population resequencing of global isolate collections of the rice blast fungus. PlasmoDB added whole-genome sequences of *Plasmodium* isolates collected on the Thai–Myanmar border between 2001 and 2014, ToxoDB added SNPs for an apicomplexan enteric pathogen, and CryptoDB added United Kingdom *Cryptosporidium* isolate collections. Across the resource, SNP records are now represented

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

as Short Variants, covering both SNPs and indels for a more complete view of genetic variation. Functional-genomic additions include two piggyBac essentiality screens in *Plasmodium knowlesi*, a genome-wide CRISPR screen in *Toxoplasma*, and updated phenotype records from PHI-base and the Candida Genome Database. ToxoDB also added a restriction fragment length polymorphism (RFLP) genotyping resource compiled from multiple studies.

Community-curated and pre-publication resources

VEuPathDB also continues to provide valuable resources that incorporate community expertise, consolidate data generated across the research community, or make data available ahead of formal publication. The updated *Trypanosoma brucei* TREU927 structural annotation in TriTrypDB revised untranslated regions (UTRs) through community Apollo curation, and AmoebaDB released a pre-publication genome-browser annotation track for *Acanthamoeba castellanii*. PlasmoDB provided early access to the *P. knowlesi* piggyBac essentiality screens and the hyperLOPIT proteome noted above. CryptoDB added a community gp60 subtype resource (https://cryptodb.org/cryptodb/app/static-content/gp60.html) [11], and ToxoDB consolidated published RFLP genotype data from multiple studies into a single searchable resource to which users can also submit new RFLP genotypes.

Connecting pathogen and host data

Across these additions, a recurring theme is the ability to examine pathogen biology alongside the biology of hosts and vectors. Host-response transcriptomes and the *Cryptosporidium parvum*-infected HCT-8 membrane proteome data in HostDB pair directly with pathogen-side data in CryptoDB and FungiDB; the dual transcriptomes from Gambian malaria patients and the paired *Plasmodium* and *Anopheles* single-cell datasets capture both sides of infection and transmission in a single study; new vector and intermediate-host genomes supply the genomic context needed to interpret them. Together these resources make it possible to move beyond individual datasets and examine biological processes across the pathogen, host and transmission cycle.

Improved functional annotation: gene product descriptions

This section describes improvements to functional annotation, specifically the gene product descriptions assigned to protein–coding genes, rather than structural annotation such as gene models and UTRs. VEuPathDB integrates gene and protein annotation information from multiple sources including public data repositories (e.g. GenBank, Ensembl), organism specific reference databases (e.g. Saccharomyces Genome Database), and individual research groups. The VEuPathDB gene product description workflow prioritizes expert manual curation over automated annotation pipelines and allows curators to incorporate information from recent publications and community input. To improve the consistency and informativeness of gene product descriptions, VEuPathDB uses a custom script (https://raw.githubusercontent.com/VEuPathDB/ApiCommonModel/refs/heads/master/Model/lib/psql/webready/orgSpecific/GeneProduct_p.psql) to identify uninformative descriptions, including terms such as "hypothetical protein" and "unspecified protein," as well as descriptions consisting only of locus or contig identifiers. For genes lacking an informative description from any available source, VEuPathDB applies the UniProt Annotation Rule-Based Annotation (ARBA) system to generate descriptions based on InterPro domains and taxonomic information. These rules can provide functional information including biological function, catalytic activity, pathway membership, subcellular localization, and protein domain composition. Where multiple annotation sources are available, manually curated descriptions are retained in preference to automatically generated annotations, and outdated automatic UniProt descriptions are removed where appropriate.

The resulting workflow provides a scalable approach for improving and maintaining gene product descriptions across the diverse organisms represented in VEuPathDB. Because the process incorporates current InterProScan and ARBA annotations, descriptions can be systematically updated as these resources are revised and can be generated consistently for newly integrated organisms. In the most recent update, this process resulted in revised gene product descriptions for >2.3 million genes, representing approximately one quarter of all genes in VEuPathDB. For transparency and reproducibility, the updated descriptions and previous descriptions are provided in the organism-specific download directories.

Updates to OrthoMCL

OrthoMCL is a comparative genomics resource within VEuPathDB that establishes orthology relationships across represented organisms [12, 13]. Orthology groups are built in two stages—all-versus-all similarity searching and clustering across a taxonomically diverse set of Core organisms, followed by addition of Peripheral organisms—and in the latest release (OrthoMCL 8, August 2026), protein sequences were aligned with VEuPathDB genome data, yielding >9.2 million proteins from 800 species clustered into over 819 000 orthology groups. Recent improvements, including DIAMOND [14] for similarity searches and OrthoFinder2 [15] for clustering, have produced more inclusive groups spanning more distantly related taxa, while new features—interactive phylogenetic trees showing gene duplication and diversification, a Similar Groups tool for identifying related protein families, and a Map Proteins to OrthoMCL tool for assigning orthology groups to newly sequenced or metagenomic sequences—have expanded its utility. These relationships are integrated throughout VEuPathDB, appearing on gene pages and underpinning synteny analyses, imputed Enzyme Commission numbers, and Search Strategy functions for transforming proteins to orthologs in other organisms or defining phyletic pattern of proteins of interest, positioning OrthoMCL as a resource connecting VEuPathDB's organismal datasets and letting users move between genes, protein families, evolutionary relationships, and function.

AI-assisted data analysis and curation

VEuPathDB incorporates artificial intelligence where appropriate and useful to facilitate annotation, data interpretation, and dataset curation.

AI-assisted gene expression analysis

An experimental AI Expression Summary feature uses Claude to analyze the numerical data and metadata associated with individual transcriptomics experiments, generating concise summaries and an estimated biological relevance score (Fig. 1,

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

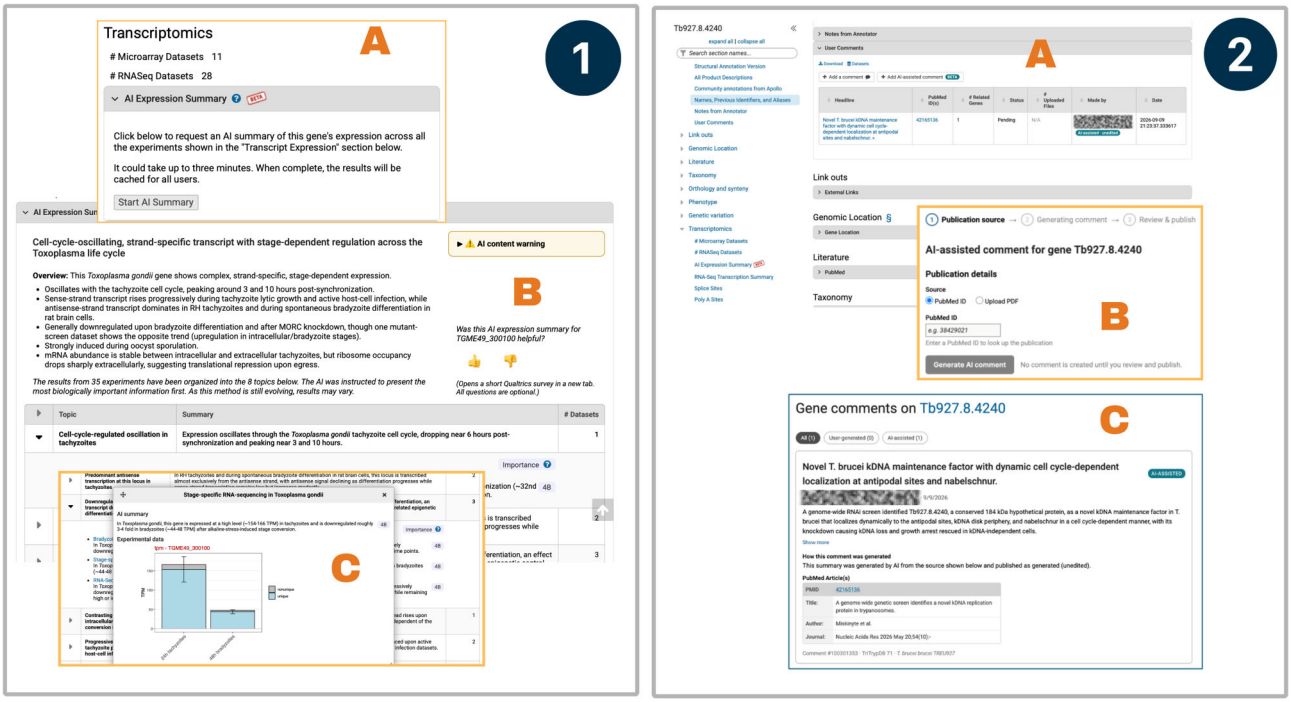


Figure 1. AI-assisted user comments and data summarization. Panel (1): AI summarization of transcriptomics data is initiated on demand (**A**) and summaries appear as they are generated and cached to conserve on resources (**B**). Each AI assertion can be corroborated by exploring the underlying data (**C**). Panel (2): Community driven AI-assisted gene annotation. User comments can be incorporated on gene pages (**A**) and users can utilize AI to summarize gene information from open access publications or uploaded PDFs (**B**). Published summaries immediately appear on gene pages and are owned by the user who can edit or delete them in the future (**C**). "Note, username has been blurred in panel 2A and C."

panel 1). To conserve AI usage, summaries are generated only when a user requests them, and all summaries are cached for future access [Fig. 1, panel 1(A)]. Summaries across experiments are subsequently grouped to identify recurring expression patterns, providing researchers with an accessible overview of the available expression evidence directly from gene pages [Fig. 1, panel 1(B)]. Specific AI assertions can be corroborated by examining graphical representations of each experiment [Fig. 1, panel 1(C)]. The feature is currently available for genes represented in at least five transcriptomics datasets and is explicitly presented as experimental, with user feedback being collected to assess its utility and guide further improvement.

Community-driven AI-assisted gene annotation

A new User-Comment interface provides a streamlined mechanism for researchers to contribute evidence and update information associated with individual genes [Fig. 1, panel 2(A)]. The interface incorporates AI-assisted drafting, allowing users to provide a PubMed identifier or upload a publication PDF and automatically generate a structured annotation from the source material [Fig. 1, panel 2(B)]. Users can review and edit the AI-generated content before publication, providing a simple workflow for sharing recent findings, highlighting relevant publications, and keeping gene-level information up to date [Fig. 1, panel 2(C)]. Published user comments become searchable and are made available on the production VEuPathDB site with minimal delay, enabling community-contributed knowledge to be rapidly incorporated into the resource and made accessible to other researchers.

AI-assisted dataset curation

AI is also being applied to the internal process of integrating new datasets into VEuPathDB. This system employs specialized AI tools to guide dataset curators through established, dataset-specific workflows, including metadata retrieval and interpretation, source data processing and transformation, and creation or adjustment of configuration files for VEuPathDB data loading. The AI skills are organism-agnostic and function as an assistant to the outreach team, taking advantage of established internal approaches, providing options for manual curation and oversight (https://github.com/VEuPathDB/dataset-curator). Different skills are used for various dataset types (e.g. genomes, RNA-Seq data), with incremental instructions and validation resources provided as needed. Initial support includes genome assembly and bulk RNA-seq datasets, with plans for additional types. This approach aims to minimize manual repetition, boost consistency and productivity, and decrease curation errors, all while maintaining human expert oversight and reproducibility during dataset integration.

"My Datasets": user-driven data integration

Traditional staff-led integration remains a core strength of VEuPathDB, providing carefully curated, annotated, and quality-controlled datasets that support consistent and reliable data discovery and analysis. However, the expertise and infrastructure required for this level of curation naturally limit the number of datasets that can be integrated at any given time and may extend the timeline from data generation to dissemination. To complement this established model and expand our capacity to make data available more rapidly, we de-

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

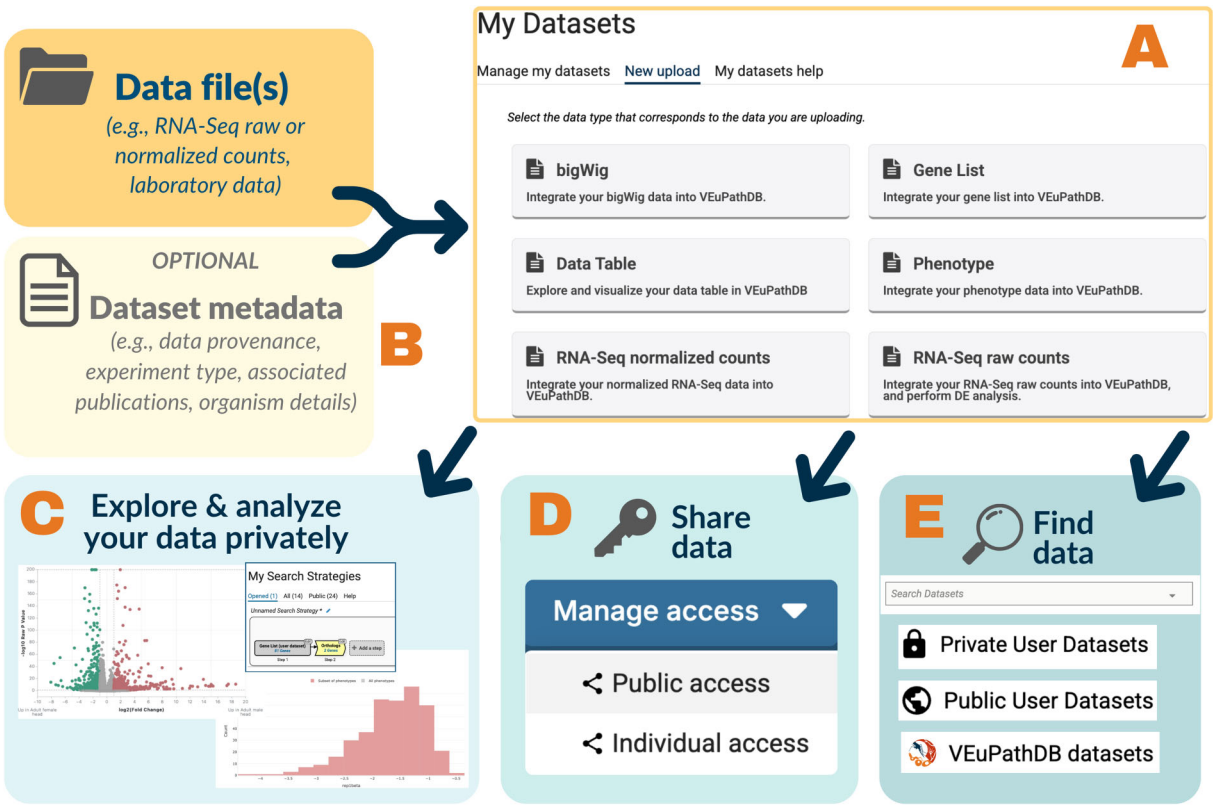


Figure 2. "My Datasets" workspace on VEuPathDB, illustrating workflow. The "My Datasets" workspace can accommodate user-uploaded data of several types (**A**) and associated dataset-level metadata (**B**). Once installed, users can explore and analyze their data using interactive tools (**C**). Providers can keep their user datasets private, share them with specific individuals, or make them public (**D**). Integrated user datasets can be searched alongside traditionally integrated datasets in the Datasets table (**E**).

veloped the "My Datasets" workspace (Fig. 2), a user-driven framework that places dataset upload, management, and sharing directly in the hands of data providers. Once integrated, user datasets can be explored using VEuPathDB's web-based search, visualization, and analysis tools, allowing researchers to analyze newly generated data in the context of existing genomic and functional datasets without requiring local software installation or custom programming.

The "My Datasets" workspace (https://veupathdb.org/veupathdb/app/workspace/datasets/new) currently supports several data types, including gene lists, RNA-Seq expression data (raw counts or normalized expression values), phenotype data, general data tables, and bigWig files. During upload, users provide metadata that establish the context and provenance of the dataset, including a description, names of investigators and collaborators, associated publications, and organisms represented. For phenotype, data table, and raw-count RNA-Seq datasets, users can also provide variable annotations such as labels and definitions. Integration links the uploaded data to the appropriate organisms and genomic features represented in VEuPathDB, enabling their use with existing genomic analysis tools.

Data providers retain control over access to their user datasets. Newly integrated datasets are private by default and accessible only to the provider. Providers may grant individual access to specific collaborators, who can then view, explore, and download the private dataset, or make the dataset public, allowing it to be discovered, explored, and downloaded by the broader research community. This provider-controlled access model enables datasets to be shared with collaborators immediately and made public when appropriate, supporting earlier dissemination of research data.

User datasets are cataloged and summarized alongside datasets integrated through VEuPathDB's traditional staff-led workflow in the "Datasets" table. This table is personalized according to dataset access: in addition to VEuPathDB-curated datasets and public user datasets, authenticated users can find their own private datasets and private datasets shared with them by collaborators. Private datasets belonging to other users remain undiscoverable.

Public sharing through "My Datasets" provides a low-cost mechanism for researchers to meet data-sharing requirements of funders and journals, which commonly ask that data be publicly accessible, documented with metadata, and citable, while supporting long-term access and reuse. Once made public, a user dataset can be discovered, explored, and downloaded by any researcher; it is accompanied by provider-supplied metadata describing the dataset, its investigators, associated publications, and organisms; and it receives a persistent identifier that can be cited in publications and data management plans. This identifier allows each user dataset to be cited and shared, supporting attribution, and scientific credit for data providers while promoting data reuse and reproducible research.

For supported data types, user datasets are integrated directly with VEuPathDB's flexible "Search Strategy" system.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

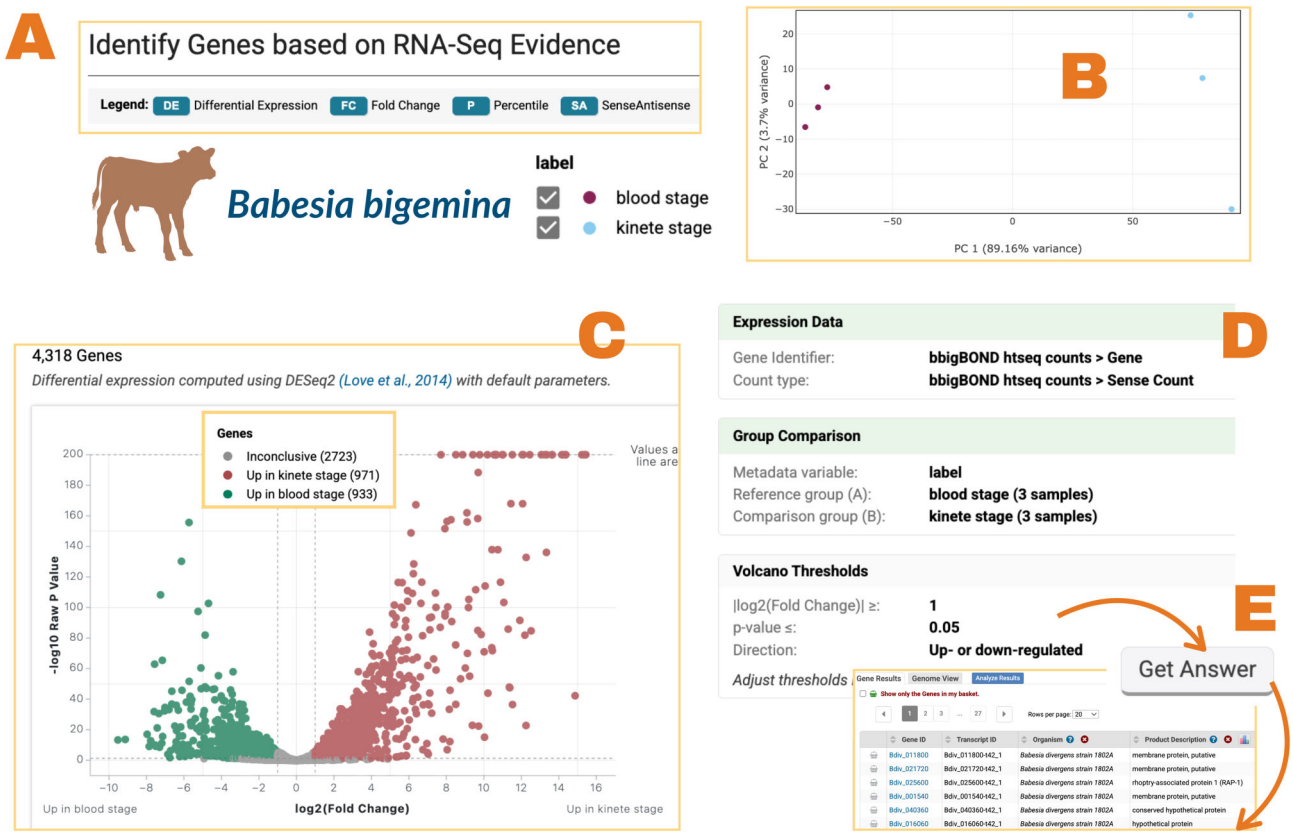


Figure 3. Transcriptomics differential expression analysis interface on VEuPathDB. The interface supports differential expression analysis among a suite of tools for identifying genes based on RNA-Seq evidence (**A**). In this example, differential expression of *B. bigemina* genes was evaluated between the calf blood and kinete stages [16]. A PCA plot provides an overview of variation among samples (**B**), while a volcano plot visualizes differential gene expression between the two stages (**C**). Effect-size and *P*-value thresholds can be adjusted (**D**) to highlight genes meeting user-defined criteria for retrieval (**E**).

For example, researchers can identify genes associated with a molecular phenotype from a phenotype dataset or genes meeting specified expression criteria in an RNA-Seq dataset. Gene sets identified from user datasets can then be transformed or combined with results from searches of VEuPathDB-curated genomic and functional data. This allows providers to analyze their own experimental results in the context of extensive staff-curated genomic data using VEuPathDB's interactive, web-based bioinformatics tools.

Data table and phenotype user datasets can additionally be explored in the interactive "Dataset Explorer" workspace. Dataset metadata populates a Dataset Details page documenting the dataset's scientific context, provenance, and associated investigators and publications. The Browse & Subset interface provides an overview of dataset variables, including their definitions, data completeness, distributions, and summary statistics, allowing researchers to rapidly assess the dataset's content and suitability for further analysis. Analytical tools embedded in the "Dataset Explorer" enable users to interactively subset data by applying filters in "Browse & Subset," generate volcano plots, principal component analysis (PCA) plots, histograms, and summary tables using point-and-click graphing tools in the Visualize interface, and explore geolocated data on the "Interactive Map." Plots can be downloaded for use in presentations and publications, and raw data can also be downloaded for offline analysis and visualization.

The "My Datasets" workspace on VEuPathDB's organism-specific component sites is designed primarily for data types that integrate with genomic resources and analysis tools for supported organisms. Clinical, epidemiological, field, population biology, and other tabular datasets that do not require genomic integration can instead be uploaded to dataExplorer (dataExplorer.org), a complementary platform for ingesting, exploring, and sharing structured data from any domain.

Interactive guided interface for differential expression analysis

VEuPathDB provides an interactive guided interface for differential expression analysis that guides users through the analysis workflow while allowing them to explore intermediate and final results. Users can select an RNA-Seq dataset, define experimental groups and contrasts, and configure analysis parameters within a reproducible, stepwise workflow (Fig. 3). The interface presents quality-control and exploratory visualizations alongside differential expression results, enabling users to assess patterns in their data and iteratively refine their analysis. Results can be further explored using VEuPathDB's integrated data-mining and visualization tools, facilitating interpretation of differentially expressed genes in the context of functional annotations and other available datasets. This approach combines the flexibility of an interactive analysis environment with the convenience of a guided workflow, making differential expression analysis accessible to users without requiring them to write code.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

Modernized infrastructure & workflows

As part of our modernization efforts, we have made significant improvements to the infrastructure supporting VEuPathDB data integration and delivery. We have migrated our databases from Oracle to PostgreSQL, eliminating costly licensing restrictions while enabling us to leverage advanced database technologies, including table partitioning and bidirectional logical replication. This transition provides a more scalable and sustainable architecture, allowing >800 organisms to be managed within a single unified database rather than across 15 siloed databases. In addition to reducing infrastructure and maintenance costs, this architecture creates new opportunities for more efficient data loading and integration across organisms.

We have also modernized our data processing architecture. Our genome and proteome workflows have been migrated from an inaccessible in-house system to Nextflow, with workflows now publicly available on GitHub, using pre-existing nf-core tools where relevant. All software tools used in these workflows are containerized using Docker, using pre-existing publicly available Docker containers when they are available, and making new custom containers publicly available in DockerHub where relevant. Together, these changes improve the reproducibility, transparency, and portability of our data-processing workflows while making them easier to maintain, scale, and extend. Collectively, the infrastructure and workflow updates provide a more flexible foundation for continued growth while improving performance, sustainability, and cost efficiency.

Conclusions and future directions

The past 2 years have tested VEuPathDB more severely than any period in its history, and the resource has come through with its data holdings, curation standards, and user community intact. The priorities ahead follow directly from the developments described above. Data integration continues at pace, with new genomes and functional genomic datasets loaded at every release, expanded coverage of fungal and oomycete plant pathogens and their hosts, and continued growth of OrthoMCL. The "My Datasets" workspace will be extended to accommodate additional data types, beginning with genetic variation data, so that a wider range of user-generated results can be integrated and analyzed alongside staff-curated data. In parallel, dataExplorer will be developed further as the route for clinical, epidemiological, field, and other tabular datasets that do not require genomic integration, broadening the range of data the resource can accept without diluting the curation standards applied to genomic data. AI-assisted curation will be extended beyond genomes and bulk RNA-seq to further dataset types, and AI-assisted features for users will be developed where they demonstrably help researchers make sense of the data rather than for their own sake, guided throughout by user feedback, curator oversight, and transparency about what has been generated automatically.

The community-supported funding model introduced in March 2025 is working, but it is not yet sufficient. Subscriptions currently cover an expanding share of operating costs, and closing the remaining gap is the central sustainability challenge of the coming period: raising awareness of the subscription program among the tens of thousands of researchers who rely on the resource, converting that reliance into group-level support, and continuing to reduce operating costs through the infrastructure and workflow changes already underway. Equitable access remains a fixed constraint, with subsidized tiers ensuring that researchers are not excluded on the basis of ability to pay. Community engagement is central to both aims. VEuPathDB will continue its training and outreach program—hands-on workshops, live demonstrations and webinars, presentations at international meetings, video tutorials on the VEuPathDB YouTube channel, and online documentation—building the analytical capacity of the research community, particularly in disease-endemic countries, while sustaining the community relationships on which the funding model depends.

Acknowledgements

We are grateful to the many colleagues who have contributed to VEuPathDB over the past two decades, including former staff and team members whose work built the foundations of the resource and whose contributions remain embedded in its data, tools, and workflows. We thank our current funders, including the Wellcome Trust and Coefficient Giving, whose support enabled the resource to be restored and sustained through the transition to a community-supported model, and we thank the philanthropic donors and subscribing groups whose contributions underpin its continued operation. We also thank the University of Pennsylvania and the University of Georgia for their institutional support of VEuPathDB's transition to the subscription model. Finally, we thank the global research community—the investigators, consortia, and data providers who generated the datasets integrated into VEuPathDB, the curators and community annotators who improve them, and the users whose feedback, comments, and continued engagement shape the direction of the resource.

Author contributions: Cristina Aurrecoechea (Resources [equal], Software [equal], Validation [equal]), Evelina Basenko (Data curation [equal], Investigation [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), Daniel Beiting (Methodology [equal], Project administration [supporting], Software [equal], Supervision [equal]), Ulrike Böhme (Data curation [equal], Writing—original draft [equal], Writing—review & editing [equal]), John Brestelli (Software [equal], Supervision [equal], Validation [equal], Visualization [equal]), Stuart Brown (Data curation [equal], Resources [equal], Writing—original draft [equal], Writing—review & editing [equal]), Kathryn Crouch (Data curation [equal], Software [equal], Writing—review & editing [equal]), Richard Demko (Data curation [equal], Resources [equal], Software [equal]), Ryan Doherty (Software [equal]), Luc Elliott (Software [equal], Writing—review & editing [equal]), Steve Fischer (Conceptualization [equal], Software [equal], Validation [equal]), Bindu Gajria (Software [equal], Validation [equal]), Omar S. Harb (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [supporting], Investigation [equal], Methodology [equal], Project administration [lead], Writing—original draft [equal], Writing—review & editing [equal]), Elizabeth Harper (Software [equal]), Danica Helb (Data curation [equal], Writing—original draft [supporting], Writing—review & editing [equal]), Sufen Hu (Data curation [equal], Resources [equal]), Andrew Jones (Funding acquisition [equal], Methodology [equal], Validation [equal]), Jessica C. Kissinger (Conceptualization [supporting], Project

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

administration [supporting]), Nupur Kittur (Data curation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Wei Li (Data curation [equal], Resources [equal]), Adriana Ludwig (Data curation [supporting], Formal analysis [supporting], Software [supporting]), Robert MacCallum (Software [equal]), Antony McCabe (Data curation [supporting], Software [supporting]), Mary Ann McDowell (Conceptualization [supporting], Project administration [supporting]), Shahram Mesdaghi (Data curation [supporting], Software [supporting]), Ahmed Muharram (Conceptualization [supporting], Software [equal]), Sadik Muzemil (Data curation [supporting], Software [supporting]), Mustafa Veysi Nural (Project administration [supporting], Software [supporting]), Freyja Redmond (Data curation [supporting], Writing—review & editing [supporting]), Daniel J. Rigden (Project administration [supporting]), David S Roos (Conceptualization [equal], Funding acquisition [lead], Project administration [lead], Supervision [equal], Writing—review & editing [equal]), Drew Spruill (Resources [supporting], Software [supporting]), Jordan Tzvetkov (Data curation [supporting], Writing—review & editing [supporting]), and Mariann Winkelman (Project administration [supporting], Writing—review & editing [supporting])

Conflict of interest

None declared.

Funding

This work was supported by the Wellcome Trust Discretionary Award no. 331329/Z/25/Z; Wellcome Trust Discovery Award no. 302692/Z/23/Z; the Biotechnology and Biological Sciences Research Council (BBSRC) UKRI744; and institutional support from the University of Pennsylvania and the University of Georgia. Funding to pay the Open Access publication charges for this article was provided by University of Pennsylvania.

Data availability

VEuPathDB (https://veupathdb.org) is a freely available, open-access resource; while login is required to use the resource, no payment is required to search, browse, analyze, or download any of the data described in this manuscript. As discussed with and approved by the NAR editorial office in advance of submission, site registration (which is free) is required to use certain personalized features and to support the voluntary subscription model described in the main text, but does not gate access to data, search tools, or analyses; this arrangement falls within the flexibility introduced by the 2026 Database Issue editorial, which notes that databases forced to charge users and to introduce necessary user registration by acute funding challenges may still be invited to submit an update by prior agreement with the editor [10]. Nucleotide and protein sequence, expression, proteomic, epigenomic, and variation data integrated into VEuPathDB originate from public archives (e.g. GenBank, ENA/SRA, PRIDE) and other public sources acknowledged on each dataset's page, with full provenance, citation, and terms of reuse provided via the Datasets table (https://veupathdb.org/veupathdb/app/search/dataset/AllDatasets/result). Data produced directly by VEuPathDB—including curated gene product descriptions, OrthoMCL orthology assignments, and processed genomic datasets—are freely downloadable in standard formats (e.g. FASTA, GFF, GFF3, tab-delimited) from the VEuPathDB download site (https://veupathdb.org/veupathdb/app/downloads) and are also accessible via the VEuPathDB REST API, without restriction or embargo. Source code for the data-processing workflows and infrastructure described in this manuscript is publicly available on GitHub (https://github.com/VEuPathDB) and Zenodo (https://doi.org/10.5281/zenodo.23212818), with associated software containers available via DockerHub. No new experimental data were generated for this manuscript.

References

1. Amos B, Aurrecoechea C, Barba M *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res* 2022;50:D898–911. https://doi.org/10.1093/nar/gkab929
2. Alvarez-Jarreta J, Amos B, Aurrecoechea C *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center in 2023. *Nucleic Acids Res* 2024;52:D808–16. https://doi.org/10.1093/nar/gkad1003
3. Harb OS, McDowell MA, Roos DS. VEuPathDB resources: a platform for free online data exploration, integration, and analysis. *Methods Mol Biol* 2024;2802:573–86. https://doi.org/10.1007/978-1-0716-3838-5_19
4. Ruhamyankaka E, Brunk BP, Dorsey G *et al.* ClinEpiDB: an open-access clinical epidemiology database resource encouraging online exploration of complex studies. *Gates Open Res* 2019;3:1661. https://doi.org/10.12688/gatesopenres.13087.1
5. Oliveira FS, Brestelli J, Cade S *et al.* MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments. *Nucleic Acids Res* 2018;46:D684–91. https://doi.org/10.1093/nar/gkx1027
6. Wadman M. NIH ends funding for key parasitology database. *Science* 2024;385:1262–3. https://doi.org/10.1126/science.adt2250
7. Fernandez-Prada C, Moretti NS, do Monte-Neto RL. Critical loss: the effects of VEuPathDB defunding on global health. *Lancet Microbe* 2025;6:100980. https://doi.org/10.1016/j.lanmic.2024.100980
8. Christophides GK. Essential vector-disease resource faces shutdown without funding. *Nature* 2024;634:33. https://doi.org/10.1038/d41586-024-03196-6
9. Little T, Will a backbone of pathogen research be rescued after funders pull the plug?. *BugBitten (BMC blog)*, 2024, https://blogs.biomedcentral.com/bugbitten/2024/09/20/will-a-backbone-of-pathogen-research-be-rescued-after-funders-pull-the-plug/ (14 November 2026, date last accessed).
10. Rigden DJ, Fernández XM. The 2026 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res* 2026;54:D1–9. https://doi.org/10.1093/nar/gkaf1427
11. Robinson G, Chalmers RM, Elwin K *et al.* Deciphering a cryptic minefield: a guide to Cryptosporidium gp60 subtyping. *Curr Res Parasitol Vector Borne Dis* 2025;7:100257. https://doi.org/10.1016/j.crpvbd.2025.100257
12. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;13:2178–89. https://doi.org/10.1101/gr.1224503
13. Chen F, Mackey AJ, Stoeckert CJ *et al.* OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 2006;34:D363–8. https://doi.org/10.1093/nar/gkj123

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

14. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60. https://doi.org/10.1038/nmeth.3176
15. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;20:238. https://doi.org/10.1186/s13059-019-1832-y
16. Capelli-Peixoto J, Saelao P, Johnson WC *et al.* Comparison of high throughput RNA sequences between *Babesia bigemina* and *Babesia bovis* revealed consistent differential gene expression that is required for the *Babesia* life cycle in the vertebrate and invertebrate hosts. *Front Cell Infect Microbiol* 2022;12:1093338. https://doi.org/10.3389/fcimb.2022.1093338

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1245/9111118 by guest on 26 November 2026

Received: September 15, 2026. **Revised:** October 26, 2026. **Accepted:** October 27, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.