



HAL
open science

ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods

Patrice Baa-Puyoulet, Nicolas Parisot, Gérard Febvay, Jaime Huerta-Cepas, Augusto F Vellozo, Toni Gabaldon, Federica Calevro, Hubert Charles, Stefano Colella

► To cite this version:

Patrice Baa-Puyoulet, Nicolas Parisot, Gérard Febvay, Jaime Huerta-Cepas, Augusto F Vellozo, et al.. ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods. Database - The journal of Biological Databases and Curation, 2016, pp.1-9. 10.1093/database/baw081 . hal-01352558

HAL Id: hal-01352558

<https://inria.hal.science/hal-01352558>

Submitted on 8 Aug 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



Database tool

ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods

Patrice Baa-Puyoulet¹, Nicolas Parisot¹, Gérard Febvay¹,
Jaime Huerta-Cepas², Augusto F. Vellozo³, Toni Gabaldón^{2,4,5},
Federica Calevro¹, Hubert Charles¹ and Stefano Colella^{1*}

¹Univ Lyon, INSA-Lyon, INRA, BF2I, UMR0203, F-69621, Villeurbanne, France, ²Centre for Genomic Regulation (CRG), the Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain, ³Univ Lyon, Univ Lyon1, CNRS, LBBE, UMR5558, F-69622, Villeurbanne, France, ⁴Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain and ⁵Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, Barcelona 08010, Spain

*Corresponding author: Tel: +33 4 72 43 84 76; Fax: +33 4 72 43 85 34; Email: stefano.colella@lyon.inra.fr

Present address: Jaime Huerta-Cepas, Structural and Computational Biology Unit, EMBL Heidelberg, D-69117 Heidelberg, Germany.

Citation details: Baa-Puyoulet,P., Parisot,N., Febvay,G. *et al.* ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods. *Database* (2016) Vol. 2016: article ID baw081; doi:10.1093/database/baw081

Received 1 December 2015; Revised 31 March 2016; Accepted 25 April 2016

Abstract

Arthropods interact with humans at different levels with highly beneficial roles (e.g. as pollinators), as well as with a negative impact for example as vectors of human or animal diseases, or as agricultural pests. Several arthropod genomes are available at present and many others will be sequenced in the near future in the context of the i5K initiative, offering opportunities for reconstructing, modelling and comparing their metabolic networks. In-depth analysis of these genomic data through metabolism reconstruction is expected to contribute to a better understanding of the biology of arthropods, thereby allowing the development of new strategies to control harmful species. In this context, we present here ArthropodaCyc, a dedicated BioCyc collection of databases using the Cyc annotation database system (CycADS), allowing researchers to perform reliable metabolism comparisons of fully sequenced arthropods genomes. Since the annotation quality is a key factor when performing such global genome comparisons, all proteins from the genomes included in the ArthropodaCyc database were re-annotated using several annotation tools and orthology information. All functional/domain annotation results and their sources were integrated in the databases for user access. Currently, ArthropodaCyc offers a centralized repository of metabolic pathways, protein sequence

domains, Gene Ontology annotations as well as evolutionary information for 28 arthropod species. Such database collection allows metabolism analysis both with integrated tools and through extraction of data in formats suitable for systems biology studies.

Database URL: <http://arthropodacyc.cycadsys.org/>

Introduction

More than 10 years have passed since the publication of the initial sequencing and analysis of the human genome (1, 2), which has had a great impact on the way we investigate biological processes, notably with the development of novel technologies enabling comprehensive genomic analyses (3). The genomes of several other organisms were sequenced before and after the human genome, starting with the animal models: *Drosophila melanogaster* (4), *Mus musculus* (5) and *Rattus norvegicus* (6). These data are driving the development of genomics-based research approaches to study the biology of many living organisms beyond humans and established model organisms. Recently, a large community of researchers have launched the Genome 10K project to obtain the full genome sequences of 10 000 vertebrate species (7). A similar initiative has been launched by the Arthropod Genomics consortium: the i5K initiative, which aims at sequencing the genomes of 5000 arthropod species (<http://arthropodgenomes.org/wiki/i5K>) (8, 9).

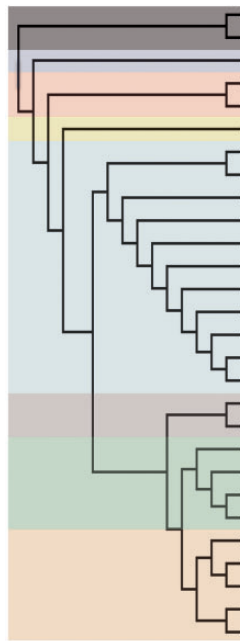
The availability of the full genome sequence of an organism allows researchers to have a complete view of its metabolism. The BioCyc collection of Pathway/Genome Databases (PGDBs) (10) constitutes a key resource for studying the metabolism of multiple organisms as it enables comparative studies. The first database of the collection, EcoCyc (11), is at present a comprehensive resource to study *Escherichia coli* biology (12). The quality of these databases is strongly linked to the annotation used to generate them and, in the most recent release (June 24, 2015—version 19.1), only seven databases are intensively manually curated and frequently updated (BioCyc Tier 1 PGDBs): EcoCyc (12), MetaCyc (13), HumanCyc (14), AraCyc (15), YeastCyc (16), LeishCyc (17) and TrypanoCyc (18). Such expert driven annotation is only possible for large communities of scientists working on the same model and, consequently, the majority of the 5711 BioCyc PGDBs available in this release are computationally derived: 39 are subject to moderate manual curation (BioCyc Tier 2 PGDBs) and 5455 to no manual curation at all (BioCyc Tier 3 PGDBs) (see <http://biocyc.org/biocyc-pgdb-list.shtml> for an updated listing). The upcoming deluge of fully sequenced genomes, driven by NGS technology, demands the development of a novel genomic

infrastructure (19). To contribute to the need of standardized automated annotation, we developed a Cyc Annotation Database System (CycADS) (20). CycADS was successfully used to generate AcypiCyc (<http://acypicyc.cycadsys.org>), a database dedicated to the pea aphid *Acyrtosiphon pisum* metabolism that was developed during the annotation phase of the genome of this insect (21). As many other arthropod genomes are available, and many more will be in the future in the context of the i5K initiative, we decided to develop a collection of BioCyc Tier 3 pathway/genome databases for arthropods using the uniform and enriched automated functional annotations provided by the CycADS system.

Implementation

CycADS annotation management system

ArthropodaCyc is a collection of BioCyc PGDBs that contains, at the time of writing (March 2016), the metabolic network of 28 arthropods with sequenced genomes (4, 21–45), including 25 insects, two arachnids and a branchiopod [Figure 1, note that phylogenetic relationships between species are displayed using a cladogram based on available data (46–50)]. All databases in ArthropodaCyc were generated using CycADS (20), an annotation management system programmed in Java (Model-View-Controller structure) and SQL that was originally developed for the annotation of the pea aphid genome (21). CycADS facilitates the collection and management of information obtained from both genomic data and different protein annotation methods in a SQL database. A pipeline to filter for bacterial contaminants was developed and integrated in the protein functional annotation system, which involves multiple methods (see below for a detailed description). All data collected in CycADS were then extracted and formatted to generate, for each organism, an *ad hoc* input file (a BioCyc ‘Path-o-logic file’ format) used by the Pathway Tools software (51) to produce BioCyc-like enriched metabolic database (20) (Supplementary Figure S1). Genomes are included in ArthropodaCyc if they fulfil the following criteria: (i) the genome sequence is published and (ii) the sequence data can be downloaded in appropriate formats (comprehensive GFF or Genbank file with compatible gene/mRNA/protein features). Each



Class	Order	Species	Database name	EC Category						Total reactions with full or partial EC numbers	REFs.
				- 1 - Oxidoreductases	- 2 - Transferases	- 3 - Hydrolases	- 4 - Lyases	- 5 - Isomerases	- 6 - Ligases		
Arachnida	Ixodida	<i>I. scapularis</i>	IxoscCyc	604 (28%)	745 (34%)	496 (23%)	176 (8%)	56 (3%)	101 (5%)	2178	(60)
Arachnida	Trombidiformes	<i>T. urticae</i>	TeturCyc	523 (26%)	727 (36%)	450 (23%)	143 (7%)	62 (3%)	88 (4%)	1993	(22,57)
Branchiopoda	Cladocera	<i>D. pulex</i>	DappuCyc	618 (28%)	779 (35%)	533 (24%)	148 (7%)	67 (3%)	97 (4%)	2242	(23,58)
Insecta	Hemiptera	<i>R. prolixus</i>	RhoprCyc	562 (27%)	772 (36%)	489 (23%)	139 (7%)	59 (3%)	99 (5%)	2120	(39,60)
Insecta	Hemiptera	<i>A. pisum</i>	AcypiCyc	546 (25%)	787 (36%)	498 (23%)	172 (8%)	59 (3%)	95 (4%)	2157	(21,52)
Insecta	Phthiraptera	<i>P. humanus corporis</i>	PedhuCyc	525 (25%)	765 (37%)	474 (23%)	177 (8%)	54 (3%)	99 (5%)	2094	(24,60)
Insecta	Hymenoptera	<i>C. solmsi</i>	CersoCyc	545 (27%)	736 (36%)	463 (23%)	144 (7%)	56 (3%)	103 (5%)	2047	(42,61)
Insecta	Hymenoptera	<i>N. vitripennis</i>	NasviCyc	547 (26%)	737 (35%)	495 (24%)	144 (7%)	61 (3%)	98 (5%)	2082	(25,55)
Insecta	Hymenoptera	<i>A. mellifera</i>	ApimeCyc	549 (26%)	783 (37%)	492 (23%)	152 (7%)	56 (3%)	108 (5%)	2140	(26,55)
Insecta	Hymenoptera	<i>L. albipes</i>	LasalCyc	517 (25%)	740 (36%)	474 (23%)	180 (9%)	51 (2%)	101 (5%)	2063	(43,55)
Insecta	Hymenoptera	<i>H. saltator</i>	HarsaCyc	563 (26%)	752 (35%)	491 (23%)	178 (8%)	52 (2%)	105 (5%)	2141	(27,55)
Insecta	Hymenoptera	<i>L. humile</i>	LinhuCyc	537 (26%)	752 (36%)	480 (23%)	145 (7%)	54 (3%)	98 (5%)	2066	(55)
Insecta	Hymenoptera	<i>C. floridanus</i>	CamflCyc	544 (26%)	767 (36%)	485 (23%)	177 (8%)	53 (2%)	101 (5%)	2127	(27,55)
Insecta	Hymenoptera	<i>P. barbatus</i>	PogbaCyc	553 (26%)	766 (36%)	469 (22%)	174 (8%)	52 (2%)	98 (5%)	2112	(28,55)
Insecta	Hymenoptera	<i>S. invicta</i>	SolinCyc	517 (26%)	690 (35%)	448 (23%)	169 (9%)	52 (3%)	95 (5%)	1971	(29,55)
Insecta	Hymenoptera	<i>A. cephalotes</i>	AttcecCyc	555 (26%)	752 (36%)	461 (22%)	173 (8%)	51 (2%)	104 (5%)	2096	(30,55)
Insecta	Hymenoptera	<i>A. echinator</i>	AcrecCyc	545 (26%)	757 (37%)	469 (23%)	143 (7%)	52 (3%)	105 (5%)	2071	(31,55)
Insecta	Coleoptera	<i>D. ponderosae</i>	DenpoCyc	517 (25%)	730 (36%)	495 (24%)	151 (7%)	50 (2%)	106 (5%)	2049	(40)
Insecta	Coleoptera	<i>T. castaneum</i>	TricaCyc	560 (25%)	787 (36%)	521 (24%)	177 (8%)	55 (3%)	99 (5%)	2199	(32,54)
Insecta	Lepidoptera	<i>P. xylostella</i>	PluxyCyc	700 (28%)	846 (34%)	601 (24%)	198 (8%)	64 (3%)	116 (5%)	2525	(41)
Insecta	Lepidoptera	<i>B. mori</i>	BommoCyc	558 (26%)	746 (35%)	526 (25%)	148 (7%)	61 (3%)	106 (5%)	2145	(44,45,62)
Insecta	Lepidoptera	<i>D. plexippus</i>	DanplCyc	573 (27%)	748 (35%)	521 (24%)	143 (7%)	63 (3%)	98 (5%)	2146	(33,59)
Insecta	Lepidoptera	<i>H. melpomene</i>	HelmeCyc	538 (26%)	724 (36%)	488 (24%)	134 (7%)	53 (3%)	95 (5%)	2032	(34,53)
Insecta	Diptera	<i>A. gambiae</i> PEST	AnogaCyc	548 (26%)	759 (36%)	500 (24%)	143 (7%)	59 (3%)	99 (5%)	2108	(35,60)
Insecta	Diptera	<i>A. aegypti</i>	AedaeCyc	564 (26%)	761 (35%)	516 (24%)	142 (7%)	61 (3%)	106 (5%)	2150	(36,60)
Insecta	Diptera	<i>C. quinquefasciatus</i>	CulquCyc	577 (27%)	767 (36%)	511 (24%)	142 (7%)	60 (3%)	96 (4%)	2153	(37,60)
Insecta	Diptera	<i>D. melanogaster</i>	DromeCyc	525 (25%)	743 (36%)	495 (24%)	141 (7%)	59 (3%)	99 (5%)	2062	(4,56)
Insecta	Diptera	<i>G. morsitans</i>	GlomoCyc	512 (26%)	725 (37%)	452 (23%)	139 (7%)	53 (3%)	90 (5%)	1971	(38,60)

Figure 1. ArthropodaCyc databases list and summary. This table shows the distribution of reactions in the Cyc databases across the six top-level categories identified by the Enzyme Commission (E.C.). Included in this table are all reactions in each database which have been assigned either full or partial E.C. numbers, and for which an enzyme has been identified (these statistics do not include pathway holes). Phylogenetic relationships between species are displayed using a cladogram based on available data (46–50).

organism database summary page contains updated information relative to the genome data release used to generate it.

Filtering possible bacterial contaminations

Organisms across all kingdoms of life are associated with microbial partners, with interactions ranging from parasitism to mutualism. Arthropods are no exception and can harbour microorganisms at their external surface, or internally as endosymbionts, gut microbiota, parasites or pathogens. Despite the use of specific DNA extraction protocols, massive sequencing of arthropod genomes may generate sequences contaminated by prokaryotic DNA. Since the ArthropodaCyc databases aims at collecting high-quality functional annotations, we decided to implement a pipeline for the detection of putative contaminant bacterial sequences, to be used before the reconstruction step of the arthropod metabolic pathways.

First, genomic sequences, annotation files (GFF/GBK) and protein sets of each arthropod genome project were retrieved from public repositories (52–62). Genomic sequences smaller than 50 kbp were compared to the NCBI's

RefSeq prokaryotic genome sequences database (63) using BLASTN (64). Contaminant genomic sequences were identified using a 90% identity threshold over at least 90% of the query length. A BLASTP against the NCBI RefSeq prokaryotic protein sequences database was thus performed to check for bacterial contaminations within the remaining proteins. Protein sequences were filtered based on BLAST results using three different criteria: (i) at least 90% amino acid identity over at least 90% of the query length, (ii) at least 90% amino acid identity over at least 50% of both the query and the hit lengths or (iii) at least 95% identity over a sliding window of 100 amino acids. To reduce the risk of removing arthropod sequences (false-positives), we included a last step performing a BLASTP ($\geq 80\%$ amino acid identity over at least 80% of the query length) of the putative contaminant protein list against an invertebrate subset of the reference UniProtKB/Swiss-Prot protein sequence database (65). Proteins with positive hits were reintegrated in the annotation process. Lists of putative bacterial contaminants detected within the 28 arthropod genomes of ArthropodaCyc are provided in [Supplementary Table S1](#). The proteins identified as putative bacterial contaminants were flagged upon extraction and they were not

used in the Pathway Tools reactions inference (see below). However, they still appear in the database with a gene/protein page and the information about their status of contaminants.

The functional annotation pipeline

We used multiple methods to perform a functional annotation: the online KAAS-KEGG annotation pipeline (66) and the PRIAM (67), Blast2GO (68, 69) and InterProScan (70) pipelines with a local installation for faster data generation (summary of results in Supplementary Table S2). These methods generated functional information (EC number, KEGG Orthology and Gene Ontology) related to the protein sequences and all annotations were collected in the database using flexible annotation loaders (annotation collector module) available in CycADS. Default parameters were used for software configurations and the BLAST alignments (prior to the Blast2GO analysis) were performed against the reference UniProtKB/Swiss-Prot protein sequences database (65). All annotation data were extracted from the CycADS SQL database for each analysed genome and collected in a Pathologic file that was used in Pathway Tools (51) to generate the corresponding BioCyc PGDB.

For several arthropods, a genome wide phylogenetic analysis performed using the Phylome pipeline and collected in PhylomeDB (71) was available. In those cases, Gene Ontology annotations were transferred using orthology relationships to *Drosophila melanogaster* and integrated in the ArthropodaCyc databases using CycADS as previously described (20). Furthermore, for the arthropods with a Phylome analysis, the orthology predictions generated in MetaPhOrs (June 2015 release) (72) using a combination of phylogenetic information derived from different databases were included into the corresponding BioCyc databases using the orthology functionality of Pathway Tools.

Using CycADS, enriched gene records were automatically generated in the Pathologic file format imported by Pathway Tools. Each Pathologic file record contained the gene and the gene product names, synonyms, sequence structural information, as well as the annotations including Enzyme evidences (E.C. numbers), KEGG Orthology, Gene Ontology and MetaPhOrs orthology that were integrated in the databases. In the note section for each gene/protein page, the information relative to the annotation results are recorded to allow the researchers to evaluate the confidence for each putative function assigned to a protein (20). Useful external cross-links, such as to NCBI's RefSeq or genomic databases of arthropod communities (i.e. AphidBase, VectorBase and Hymenoptera Genome

Database), are also integrated thanks to the CycADS pipeline. Moreover, as InterProScan (70) analysis identifies functional domains, we included links to the InterPro external database identifiers when appropriate (see Figure 2 example page).

It is important to underline that, as data formats in genomics can be very disparate depending on the source file format used (i.e. GFF, Gbk), data on each arthropod genome were manually checked to ensure that a unique identifier for genes and products was present and that clear relations could be established among the different features. The flexible CycADS parsers were parameterized using its detailed configuration file (20). Finally, all automatic tasks of Pathway Tools consistency checker were run before databases saving and publishing. With the perspective to make available an up-to-date database representative of the fast evolving field of arthropod genomics, ArthropodaCyc will be updated annually.

Discussion on features and usage

The ArthropodaCyc collection of enriched BioCyc databases for arthropods whose genome has been fully sequenced and assembled is a key resource for all members of the Arthropod Genomics Consortium (http://arthropodgenomes.org/wiki/Main_Page). Our collection takes advantage of CycADS (20): a powerful annotation management system allowing to manage multiple genomes and to generate a set of BioCyc database where each organism has been annotated using identical tools and automatized procedures. Furthermore, our collection is enriched by phylogeny-based orthology predictions available in PhylomeDB/MetaPhOrs (71, 72) and customized hyperlinks to organism specific genome browsers. The ArthropodaCyc collection of databases takes also full advantage of the rich BioCyc interface and tools for metabolism data analysis (73). Several analyses can be performed using the BioCyc online interface that includes advanced query tools (74), and powerful web-based genomic data viewers (75). Moreover, ArthropodaCyc offers the possibility to download the data in formats suitable for data analysis: either using other tools, such as for example Cytoscape (76) and MetExplore (77), or for use in personally developed analysis software and pipelines.

We are already using ArthropodaCyc to contribute to the analysis of the metabolism in genome annotation projects currently ongoing on different insect species: the green peach aphid *Myzus persicae* [manuscript in preparation], the milkweed bug *Oncopeltus fasciatus* [manuscript in preparation](78–80) and the rice weevil *Sitophilus oryzae*. Beyond the single organism analysis, the BioCyc interface provides the user with tools for comparative analyses (51)

The pathways herein are automatically generated and not manually curated, hence users should take caution when interpreting the existence or absence of metabolic pathways.

LOGIN | Why Login? | Create New Account

Enter a gene, protein, metabolite or pathway... Quick Search Gene Search

Searching *Acyrtosiphon pisum* (*Acyr_2.0*) change organism database

Home Search Genome Metabolism Analysis SmartTables Help

log in to add this to a SmartTable. ***Acyrtosiphon pisum* v2.1c (AphidBase v2.1b) Enzyme: ACYPI009936-PA**

1 Gene: **ACYPI009936** Accession Numbers: G1578-98056 (*Acypl_V21cCyc*), 7029-16245
 Synonyms: XM_001950476.2_DG2_gene, XP_001950511.1_DG2_pep, Phy00108GV_ACYPI

2 Summary:
 KO:K00861 with 2 annotation evidences using method(s):KAAS_Eukaryotes, KAAS_Genes
 GO:0008531 with 4 annotation evidences using method(s):SUPERFAMILY, Pfam, Gene3D, SMART
 GO:0009231 with 4 annotation evidences using method(s):SUPERFAMILY, Pfam, Gene3D, SMART
 GO:0005737 with a score of 2 using method(s):PhylomeDB¹
 GO:0008531 with a score of 3 using method(s):PhylomeDB¹
 GO:0009231 with a score of 3 using method(s):PhylomeDB¹
 GO:0046872 with a score of 3 using method(s):PhylomeDB¹
 GO:0005524 with a score of 3 using method(s):PhylomeDB¹
 EC:2.7.1.26 with 4 annotation evidences using method(s):Kegg-EC_Euka, Kegg-EC_Genes, Priam, Blast2GO-EC

3 Map Position: [45,708 <- 46,782] (25.78 centisomes) on Contig GL350276 Genome Browser Length: 1075 bp / 149 aa
 Molecular Weight of Polypeptide: 16.915 kD (from nucleotide sequence)

4 Unification Links: AphidBase-Gene:ACYPI009936, AphidBase-Protein:ACYPI009936-PA, AphidBase-transcript:ACYPI009936-RA, BRENDA:2.7.1.26, genomic scaffold:GL350276, InterPro:IPR015865, InterPro:IPR023465, InterPro:IPR023468, KO:K00861, PhylomeDB:Phy00108GV_ACYPI

5 Gene-Reaction Schematic:

6 GO Terms:

Biological Process:	GO:0009231 - riboflavin biosynthetic process
Molecular Function:	GO:0005524 - ATP binding GO:0008531 - riboflavin kinase activity GO:0046872 - metal ion binding
Cellular Component:	GO:0005737 - cytoplasm

Gene Class: UNCLASSIFIED

Enzymatic reaction of: **ACYPI009936-PA**
 EC Number: 2.7.1.26
 7 riboflavin + ATP <=> ADP + FMN + H⁺
 The reaction direction shown, that is, A + B <=> C + D versus C + D <=> A + B, is in accordance with the Enzyme Commission system.
 The reaction is favored in the direction shown.
 In Pathways: flavin biosynthesis IV (mammalian)

8 Gene Local Context (not to scale):

9 Exons/Introns:

Report Errors or Provide Feedback
 Please cite the following article in publications resulting from the use of *Acypl_V21cCyc*:
¹PhylomeDB scoring system: 1 for *Drosophila* transferred annotation, 2 for a one to one relation *Drosophila* transferred annotation, 3 for one to one *Drosophila* transferred annotation and conserved from ancestor
 Powered by CycADS 1.3

10 OPERATIONS
 Sequences
 Protein Sequence
 Nucleotide Sequence
 Spliced Nucleotide Sequence
 Save Nucleotide Sequence to file
 Save Protein Sequence to file
 Comparison Operations
 Show this gene in another database
 Change organisms/databases for comparison operations
 Search for this gene in other databases

Figure 2. Screenshot of an ArthropodaCyc enzyme page. The page provides several information such as: (1) gene name, accession numbers and synonym names; (2) a summary of metabolism annotation evidences from KAAS-KEGG, PRIAM, InterProScan, PhylomeDB and BLAST2GO; (3) genome position with an additional link to the corresponding genome browser, and information on gene and protein length and polypeptide molecular weight; (4) external cross-links to specific genomic databases, enzyme annotation and InterProScan domains information and to phylogeny in PhylomeDB; (5) schematics representing the reaction(s) carried out by the enzyme; (6) Gene ontology terms associated with the enzyme functions; (7) additional information on the reaction(s) carried out by the enzyme, including the pathway(s) (if any) where this reaction may occur; (8) gene local context, including neighbouring genes; (9) gene structure in terms of (added) exons/introns organization and (10) an “Operations box” offering several options for comparative analyses. Filled circles, (2) and (4), represent ArthropodaCyc specific features.

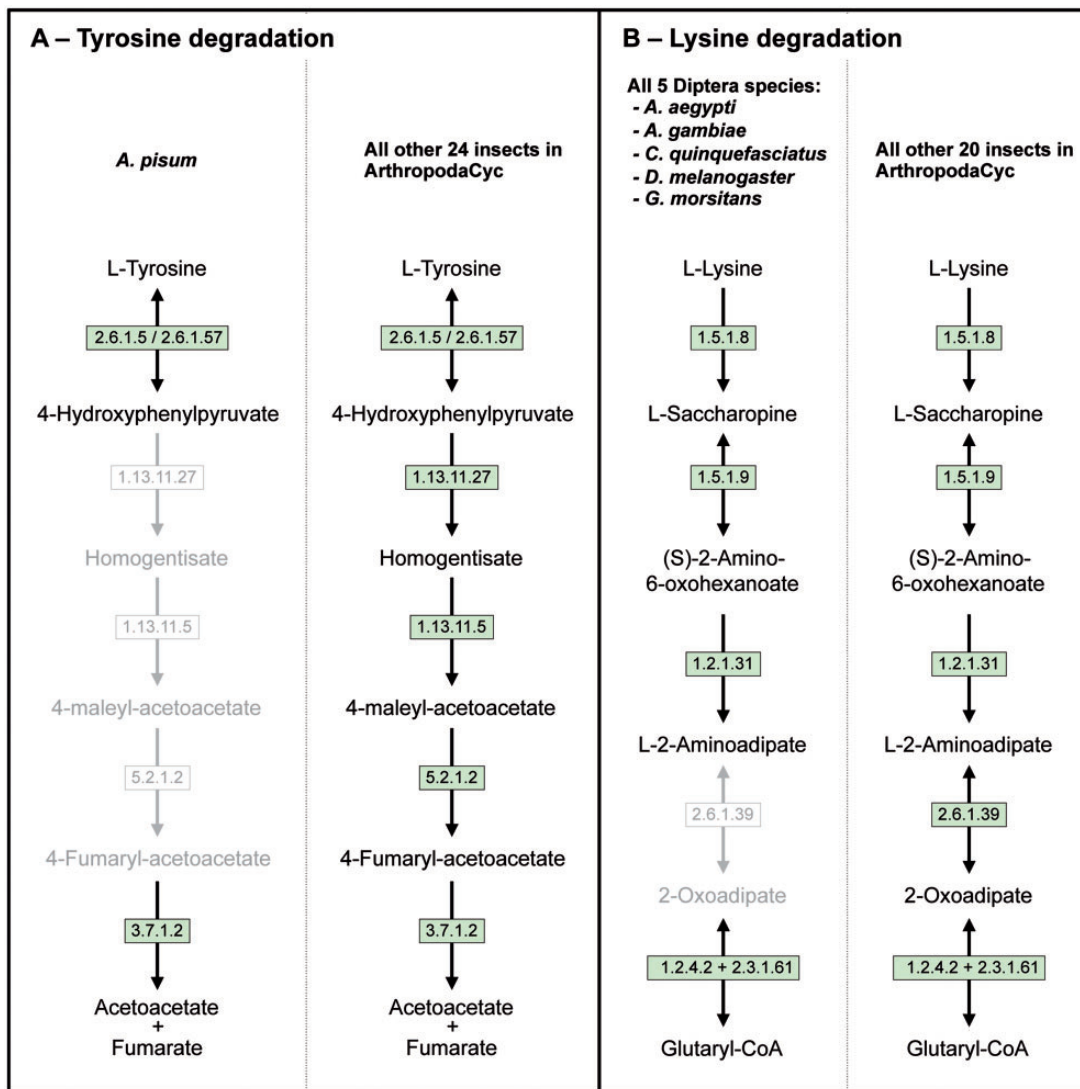


Figure 3. Two examples of insect pathway differences identified using ArthropodaCyc. (A) Pathway of tyrosine degradation, comparison between *A. pisum* and the other insects of ArthropodaCyc; (B) Pathway of lysine degradation, comparison between the five species of Diptera and the other insects of ArthropodaCyc. In each pathway, green coloured enzymes are present, while grey enzymes and reactions are absent. Enzymes: 1.13.11.5 = homogentisate 1,2-dioxygenase; 1.13.11.27 = 4-hydroxyphenylpyruvate dioxygenase; 1.2.1.31 = L-amino adipate-semialdehyde dehydrogenase; 1.2.4.2 = oxoglutarate dehydrogenase (succinyl-transferring); 1.5.1.8 = saccharopine dehydrogenase (NADP⁺, L-lysine-forming); 1.5.1.9 = saccharopine dehydrogenase (NAD⁺, L-glutamate-forming); 2.3.1.61 = dihydrolipoyllysine-residue succinyltransferase; 2.6.1.5 = tyrosine transaminase; 2.6.1.39 = 2-amino adipate transaminase; 2.6.1.57 = aromatic-amino-acid transaminase; 3.7.1.2 = fumarylacetoacetase; 5.2.1.2 = maleylacetoacetate isomerase.

to identify interesting features of a given organism metabolism that could shed light on its biology. The interest in comparative analyses will greatly grow as more arthropods are sequenced.

Even if a full comparative analysis in arthropods is beyond the purpose of this article, we provide here a few examples of its usage. ArthropodaCyc can be used for the identification of enzymes and/or pathways unique to a given organism or group of organisms. As an example, we used the present version of the ArthropodaCyc database to verify the lack of the tyrosine degradation pathway that we had originally described in the pea aphid genome by using

the AcypiCyc database (20, 21) and comparing this genome to the few insect genomes available at that time. We could confirm that *A. pisum* is the only insect lacking this pathway among the 25 available in the database at present (Figure 3A). We interpreted this loss as an explanation of the pea aphid lifestyle and we linked it to the high demand for tyrosine by pea aphids in connection with their unbalanced plant phloem sap diet. Even though the nutrition of these insects is complemented by their primary symbiont, *Buchnera aphidicola*, this bacterium provides only precursors for tyrosine biosynthesis. The enzymes catalysing the last two steps of tyrosine synthesis are in fact encoded in

the aphid genome. In particular, multiple genes coding the aspartate transaminase (E.C. 2.6.1.1), the enzyme involved in the synthesis of phenylalanine from phenylpyruvate are present in the pea aphid genome with one of them (*ACYPI004243*) specifically regulated during embryo development in aphids reproducing by parthenogenesis (81). We thus checked whether this aspartate transaminase gene expansion was also present in the genomes of the other 27 arthropods present in the ArthropodaCyc database. Remarkably, only six out of the other 27 arthropods show a gene expansion comparable to the pea aphid (5 or 4 genes), and all the other 21 arthropods present only 2 or 3 genes encoding for this enzyme (Supplementary Table S3). Even if further phylogenetic analyses would be needed to better understand the origin of the differences in this enzyme-coding gene, this is an example of another possible use of ArthropodaCyc to easily explore the number of genes and their structures for specific enzymes (an interesting genome variation beyond the relatively rare presence/absence case of enzymatic reactions).

As another example application, we also decided to explore the database to search for pathways that would be characteristic of a specific group of insects and we identified the 2-aminoadipate transaminase (E.C. 2.6.1.39) in the lysine degradation pathway as uniquely missing in the genomes of the five dipteran species available in ArthropodaCyc (*A. aegypti*, *A. gambiae*, *C. quinquefasciatus*, *D. melanogaster* and *G. morsitans*) (Figure 3B). These examples provided here show the power of ArthropodaCyc in finding differences between specific organisms that might be linked with their biology, even though for the lysine degradation it is difficult to speculate on possible reasons for this apparent loss of a complete pathway as these five dipteran species live in multiple habitats and feed on very diversified diets.

Conclusions

We present here ArthropodaCyc, the most comprehensive collection of BioCyc databases for arthropods, which we expect to be of great interest for a broad community of scientists. Several genomes of arthropods are being sequenced and many more will be sequenced in the future as part of the i5K initiative. The CycADS pipeline empowers both the development and the update of the PGDB in ArthropodaCyc. Our databases are an arthropod research resource that is also linked, whenever possible, to single organism community based genomic databases, thus offering to the researchers an integrated access to different sources of annotations.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

The authors would like to thanks: Daniel Lawson (European Bioinformatics Institute, Cambridge, UK) for the hyperlinks to/from VectorBase (<https://www.vectorbase.org/>); Fabrice Legeai and Anthony Bretaudeau at the INRA Bioinformatics Platform for Agro-ecosystems Arthropods (BIPAA) (<http://www6.inra.fr/bipaa>) for the hyperlinks to/from AphidBase. The authors would also like to thank the following resources for genome sequence data: AphidBase (<http://www.aphidbase.com/>)(52), Hymenoptera Genome Database (<http://hymenopteragenome.org/>)(55), VectorBase (<https://www.vectorbase.org/>)(60), NCBI genome (<http://www.ncbi.nlm.nih.gov/genome/>)(61), FlyBase (<http://flybase.org/>)(56), BeetleBase (<http://beetlebase.org/>)(54), OrcAE (<http://bioinformatics.psb.ugent.be/orcae/>)(57), wFleabase (<http://wfleabase.org/>)(58), MonarchBase (<http://monarchbase.umassmed.edu/>)(59), Heliconius genome project (<http://www.butterflygenome.org/>)(53), SilkDB (<http://silkworm.genomics.org.cn/>)(62), Ensembl Metazoa (<http://metazoa.ensembl.org/>), DBM-DB (<http://iae.fafu.edu.cn/DBM/>) and The TRIA project (<http://www.thetriaproject.ca/>).

Funding

The ANR Blanc Program IMetSym (ANR-13-BSV7-0016-03), the Spanish Ministry of Economy and Competitiveness grants, ‘Centro de Excelencia Severo Ochoa 2013-2017’ SEV-2012-0208 and BIO2012-37161 cofounded by European Regional Development Fund (ERDF; to T.G. group). Funding for open access charge: ANR Blanc Program IMetSym (ANR-13-BSV7-0016-03).

Conflict of interest. None declared.

References

- Lander, E.S., Linton, L.M., Birren, B. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature.*, 409, 860.
- Venter, J., Adams, M., Myers, E. *et al.* (2001) The sequence of the human genome. *Science.*, 291, 1304.
- Lander, E.S. (2011) Initial impact of the sequencing of the human genome. *Nature.*, 470, 187–197.
- Adams, M.D., Celniker, S.E., Holt, R.A. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science.*, 287, 2185–2195.
- Waterston, R.H., and Lindblad-Toh, K. Mouse Genome Sequencing Consortium. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nat. News.*, 420, 520–562.
- Gibbs, R.A., Weinstock, G.M., Metzker, M.L. *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature.*, 428, 493–521.
- Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Heredity.*, 100, 659–674.
- Robinson, G.E., Hackett, K.J., Purcell-Miramontes, M. *et al.* (2011) Creating a buzz about insect genomes. *Science.*, 331, 1386.

9. i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Heredity.*, 104, 595–600.
10. Karp,P.D., and Caspi,R. (2011) A survey of metabolic databases emphasizing the MetaCyc family. *Arch. Toxicol.*, 85, 1015–1033.
11. Karp,P.D., Riley,M., Saier,M. *et al.* (2002) The EcoCyc database. *Nucleic Acids Res.*, 30, 56–58.
12. Keseler,I.M., Collado-Vides,J., Santos-Zavaleta,A. *et al.* (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.*, 39, D583–D590.
13. Karp,P.D., Riley,M., Paley,S.M. *et al.* (2002) The MetaCyc Database. *Nucleic Acids Res.*, 30, 59–61.
14. Romero,P., Wagg,J., Green,M.L. *et al.* (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, 6, R2.
15. Mueller,L.A., Zhang,P., and Rhee,S.Y. (2003) AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol.*, 132, 453–460.
16. Caspi,R., Altman,T., Billington,R. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, 42, D459–D471.
17. Doyle,M.A., MacRae,J.I., De Souza,D.P. *et al.* (2009) LeishCyc: a biochemical pathways database for *Leishmania major*. *BMC Syst. Biol.*, 3, 57.
18. Shameer,S., Logan-Klumpler,F.J., Vinson,F. *et al.* (2015) TrypanoCyc: a community-led biochemical pathways database for *Trypanosoma brucei*. *Nucleic Acids Res.*, 42, D637–D644.
19. Parkhill,J., Birney,E., and Kersey,P. (2010) Genomic information infrastructure after the deluge. *Genome Biol.*, 11, 402.
20. Vellozo,A.F., Véron,A.S., Baa-Puyoulet,P. *et al.* (2011) CycADS: an annotation database system to ease the development and update of BioCyc databases. *Database.*, 2011, bar008–bar008.
21. International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biol.*, 8, e1000313.
22. Grbić,M., Van Leeuwen,T., Clark,R.M. *et al.* (2011) The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nat. News.*, 479, 487–492.
23. Colbourne,J.K., Pfrender,M.E., Gilbert,D. *et al.* (2011) The ecoresponsive genome of *Daphnia pulex*. *Science.*, 331, 555–561.
24. Kirkness,E.F., Haas,B.J., Sun,W. *et al.* (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Natl. Acad. Sci. USA.*, 107, 12168–12173.
25. Werren,J.H., Richards,S., Desjardins,C.A. *et al.* (2010) Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science.*, 327, 343–348.
26. Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature.*, 443, 931–949.
27. Bonasio,R., Zhang,G., Ye,C. *et al.* (2010) Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science.*, 329, 1068–1071.
28. Smith,C.R., Smith,C.D., Robertson,H.M. *et al.* (2011) Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc. Natl. Acad. Sci. USA.*, 108, 5667–5672.
29. Wurm,Y., Wang,J., Riba-Grognuz,O. *et al.* (2011) The genome of the fire ant *Solenopsis invicta*. *Proc. Natl. Acad. Sci. USA.*, 108, 5679–5684.
30. Suen,G., Teiling,C., Li,L. *et al.* (2011) The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLoS Genet.*, 7, e1002007.
31. Nygaard,S., Zhang,G., Schiött,M. *et al.* (2011) The genome of the leaf-cutting ant *Acromyrmex echinator* suggests key adaptations to advanced social life and fungus farming. *Genome Res.*, 21, 1339–1348.
32. Gibbs,R.A., Weinstock,G.M. Tribolium Genome Sequencing Consortium. *et al.* (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature.*, 452, 949–955.
33. Zhan,S., Merlin,C., Boore,J.L. *et al.* (2011) The Monarch butterfly genome yields insights into long-distance migration. *Cell.*, 147, 1171–1185.
34. Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature.*, 487, 94–98.
35. Holt,R.A., Subramanian,G.M., Halpern,A. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science.*, 298, 129–149.
36. Nene,V., Wortman,J.R., Lawson,D. *et al.* (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science.*, 316, 1718–1723.
37. Arensburger,P., Megy,K., Waterhouse,R.M. *et al.* (2010) Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science.*, 330, 86–88.
38. International Glossina Genome Initiative (2014) Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis. *Science.*, 344, 380–386.
39. Mesquita,R.D., Vionette-Amaral,R.J., Lowenberger,C. *et al.* (2015) Genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. *Proc. Natl. Acad. Sci. USA.*, 112, 14936–14941.
40. Keeling,C.I., Yuen,M.M.S., Liao,N.Y. *et al.* (2012) Draft genome of the mountain pine beetle, *Dendroctonus ponderosae Hopkins*, a major forest pest. *Genome Biol.*, 14, R27.
41. You,M., Yue,Z., He,W. *et al.* (2013) A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.*, 45, 220–225.
42. Xiao,J.H., Yue,Z., Jia,L.Y. *et al.* (2012) Obligate mutualism within a host drives the extreme specialization of a fig wasp genome. *Genome Biol.*, 14, R141.
43. Kocher,S.D., Li,C., Yang,W. *et al.* (2013) The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. *Genome Biol.*, 14, R142.
44. Xia,Q.Y., Zhou,Z.Y., Lu,C. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science.*, 306, 1937–1940.
45. Xia,Q., Wang,J., Zhou,Z. *et al.* (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect. Biochem. Mol. Biol.*, 38, 1036–1045.
46. Misof,B., Liu,S., Meusemann,K. *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science.*, 346, 763–767.
47. Biewer,M., Schlesinger,F., and Hasselmann,M. (2014) The evolutionary dynamics of major regulators for sexual

- development among Hymenoptera species. *Front. Genet.*, 6, 124–124.
48. Sieglaff,D.H., Dunn,W.A., Xie,X.S. *et al.* (2009) Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes. *Proc. Natl. Acad. Sci. USA.*, 106, 3053–3058.
 49. Sanggaard,K.W., Bechsgaard,J.S., Fang,X. *et al.* (2014) Spider genomes provide insight into composition and evolution of venom and silk. *Nat. Rev. Genet.*, 5, 3765.
 50. Ahola,V., Lehtonen,R., Somervuo,P. *et al.* (2013) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat. Rev. Genet.*, 5, 4737.
 51. Karp,P.D., Paley,S.M., Krummenacker,M. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, 11, 40–79.
 52. Gauthier,J.P., Legeai,F., Zasadzinski,A. *et al.* (2007) AphidBase: a database for aphid genomic resources. *Bioinformatics.*, 23, 783–784.
 53. Hubbard,T.J.P., Aken,B.L., Ayling,S. *et al.* (2008) Ensembl 2009. *Nucleic Acids Res.*, 37, D690–D697.
 54. Kim,H.S., Murphy,T., Xia,J. *et al.* (2009) BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. *Nucleic Acids Res.*, 38, D437–D442.
 55. Muñoz-Torres,M.C., Reese,J.T., Childers,C.P. *et al.* (2010) Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res.*, 39, D658–D662.
 56. Mcquilton,P., Pierre,S.E.S., and Thurmond,J. (2011) FlyBase 101—the basics of navigating FlyBase. *Nucleic Acids Res.*, 40, D706–D714.
 57. Sterck,L., Billiau,K., Abeel,T. *et al.* (2012) ORCAE: online resource for community annotation of eukaryotes. *Nat. Methods.*, 9, 1041–1041.
 58. Nordberg,H., Cantor,M., Dusheyko,S. *et al.* (2013) The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.*, 42, D26–D31.
 59. Zhan,S., and Reppert,S.M. (2013) MonarchBase: the monarch butterfly genome database. *Nucleic Acids Res.*, 41, D758–D763.
 60. Giraldo-Calderón,G.I., Emrich,S.J., Maccallum,R.M. *et al.* (2014) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.*, 43, D707–D713.
 61. NCBI Resource Coordinators (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 43, D6–D17.
 62. Wang,J., Xia,Q.Y., He,X.M. *et al.* (2005) SilkDB: a knowledge-base for silkworm biology and genomics. *Nucleic Acids Res.*, 33, D399–D402.
 63. Pruitt,K.D., Tatusova,T., Brown,G.R. *et al.* (2011) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, 40, D130–D135.
 64. Altschul,S.F., Gish,W., Miller,W. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–410.
 65. UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, 37, D169–D174.
 66. Moriya,Y., Itoh,M., Okuda,S. *et al.* (2007) KAAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, 35, W182–W185.
 67. Claudel-Renard,C., Chevalet,C., Faraut,T. *et al.* (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.*, 31, 6633–6639.
 68. Conesa,A., Götz,S., García-Gómez,J.M. *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.*, 21, 3674–3676.
 69. Conesa,A., and Götz,S. (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics.*, 2008, 619832.
 70. Jones,P., Binns,D., Chang,H.Y. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics.*, 30, 1236–1240.
 71. Huerta-Cepas,J., Capella-Gutierrez,S., Pryszcz,L.P. *et al.* (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, 42, D897–D902.
 72. Pryszcz,L.P., Huerta-Cepas,J., and Gabaldón,T. (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.*, 39, e32.
 73. Krummenacker,M., Paley,S., Mueller,L. *et al.* (2005) Querying and computing with BioCyc databases. *Bioinformatics.*, 21, 3454–3455.
 74. Latendresse,M., and Karp,P.D. (2010) An advanced web query interface for biological databases. *Database.*, 2010, baq006.
 75. Latendresse,M., and Karp,P.D. (2010) Web-based metabolic network visualization with a zooming user interface. *BMC Bioinformatics.*, 12, 176.
 76. Shannon,P., Markiel,A., Ozier,O. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13, 2498–2504.
 77. Cottret,L., Wildridge,D., Vinson,F. *et al.* (2010) MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res.*, 38 Suppl, W132–W137.
 78. Murali,S.C., Bandaranaike,D., Bellair,M. *et al.* *Oncopeltus fasciatus* genome assembly 1.0 | Ag Data Commons. <https://data.nal.usda.gov/dataset/oncopeltus-fasciatus-genome-assembly-10>.
 79. Vargas Jentzsch,I.M., Hughes,D.S.T., Poelchau,M., *et al.* *Oncopeltus fasciatus* Official Gene set v1.1 | Ag Data Commons. <https://data.nal.usda.gov/dataset/oncopeltus-fasciatus-official-gene-set-v11>.
 80. Hughes,D.S.T., Koelzer,S., Panfilio,K.A. *et al.* *Oncopeltus fasciatus* genome annotations v0.5.3 | Ag Data Commons. <https://data.nal.usda.gov/dataset/oncopeltus-fasciatus-genome-annotations-v053>.
 81. Rabatel,A., Febvay,G., Gaget,K. *et al.* (2013) Tyrosine pathway regulation is host-mediated in the pea aphid symbiosis during late embryonic and early larval development. *BMC Genomics.*, 14, 235.