Chapter 10

Accessing *Cryptosporidium* Omic and Isolate Data via CryptoDB.org

Susanne Warrenfeltz, Jessica C. Kissinger, and On Behalf of the EuPathDB Team

Abstract

Cryptosporidium has historically been a difficult organism to work with, and molecular genomic data for this important pathogen have typically lagged behind other prominent protist pathogens. CryptoDB (<http://cryptodb.org/>) was launched in 2004 following the appearance of draft genome sequences for both *C. parvum* and *C. hominis*. CryptoDB merged with the EuPathDB Bioinformatics Resource Center family of databases (<https://eupathdb.org>) and has been maintained and updated regularly since its establishment. These resources are freely available, are web-based, and permit users to analyze their own sequence data in the context of reference genome sequences in our user workspaces. Advances in technology have greatly facilitated *Cryptosporidium* research in the last several years greatly enhancing and extending the data and types of data available for this genus. Currently, 13 genome sequences are available for 9 species of *Cryptosporidium* as well as the distantly related *Gregarina niphandrodes* and two free-living alveolate outgroups of the Apicomplexa, *Chromera velia* and *Vitrella brassicaformis*. Recent years have seen several new genome sequences for both existing and new *Cryptosporidium* species as well as transcriptomics, proteomics, SNP, and isolate population surveys. This chapter introduces the extensive data mining and visualization capabilities of the EuPathDB software platform and introduces the data types and tools that are currently available for *Cryptosporidium*. Key features are demonstrated with *Cryptosporidium*-relevant examples and explanations.

Key words Apicomplexa, Bioinformatics, Parasite, Pathogen, Genomics, Transcriptomics, Orthology, Proteomics, SNP, Sequence analysis

1 Introduction

A member of the Eukaryotic Pathogen Bioinformatics Resource Center (<https://eupathdb.org/>) [1, 2], CryptoDB (<http://cryptodb.org/>) serves as the functional genomics database for

The original version of this chapter was revised. The correction to this chapter is available at https://doi.org/10.1007/978-1-4939-9748-0_22

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-1-4939-9748-0_10) contains supplementary material, which is available to authorized users.

Jan R. Mead and Michael J. Arrowood (eds.), *Cryptosporidium: Methods and Protocols*, Methods in Molecular Biology, vol. 2052, https://doi.org/10.1007/978-1-4939-9748-0_10, © The Author(s) 2020

Cryptosporidium and related species. CryptoDB is a free, online resource for accessing and exploring genome sequence and annotation, functional genomics data, isolate sequences, and orthology profiles across organisms. Data mining via CryptoDB requires no prior bioinformatics or computational experience since all integrated data are pre-analyzed according to standard workflows and accessed via an easy-to-use interactive web interface. CryptoDB brings a genome-wide perspective to the exploration of *Cryptosporidium* and related species by integrating the pre-analyzed data with advanced search capabilities, convenient data visualization, analysis tools, and a comprehensive record system describing genomic features and pathways.

Data mining in CryptoDB proceeds along four general paths:

1. Direct examination of record pages which compile all database information about a feature (genes, SNPs, pathways, EST, isolates, etc.) into tables and graphs.
2. An advanced search strategy system which facilitates genome-wide data mining within and across data sets, data types, and organisms (e.g., retrieve all *Cryptosporidium* orthologs that contain signal peptides or are regulated at 48 h).
3. Data visualization tools, such as a genome browser, which facilitate visualization of sequence-based data to explore gene models, single-nucleotide polymorphisms, or synteny of genomic regions across any or all integrated genome sequences.
4. Data analysis tools for functional enrichment or primary analysis via a private EuPathDB Galaxy [3, 4] instance (e.g., analysis of your RNA-seq data or variant calling).

This chapter starts with a brief overview of the database content and then describes the routine data mining methodologies outlined above. For maximum benefit, this chapter is best utilized at a computer with CryptoDB.org open in a browser. Screenshots of all relevant actions are provided as a guide for completing examples and in case offline reading is preferred.

2 Using CryptoDB

2.1 Database Content and Download

CryptoDB integrates a wide range of data types including genome sequence and annotation, whole-genome sequencing of clinical or environmental isolates, GenBank Popset [5] sequences from surveys, RNA sequence, microarray, proteomics, genome-wide RT-PCR, ESTs, and metabolic pathways. Versions of the database are released about every 2 months and incorporate new data, update existing data, and launch new features. Central to the database is the genome sequence and annotation. CryptoDB enhances these data with a series of in-house bioinformatic analyses for domain prediction, functional label assignments, and genome feature identification. In addition, functional genomics data are

Fig. 1 Homepage layout. (a) Header section. Gray menu bar (arrow 2) for quick access to all CryptoDB searches and tools. When scrolling CryptoDB pages, this menu bar remains available at the top of every page. Gene ID search (arrow 3) navigates to the gene record page of the ID entered. Gene Text Search (arrow 4) returns a list of genes whose records contain the desired text. About menu (arrow 5) with links for Help, Login, Registration, and to Contact Us via email with questions and suggestions. (b) Side panel with expandable sections containing a Data Summary (arrow 1) table listing data types available for each genome in CryptoDB, news which documents changes per release, tweets which inform on items of interest to the Crypto research community, links to other community resources, and educational materials. (c) Searches and tools section used to access all searches that return genes (arrow 6), other record types (Popset Isolates, ESTs, etc.) (arrow 7) or tools (arrow 8) for BLAST, analysis or annotation, among others. (d) Footer section with links to all EuPathDB home pages

mapped to the genome sequence for interrogation and visualization of sequence-based data in a genomic context.

The data summary table provides an overview of available data as of the submission of this chapter. Open the *Data Summary* section in the home page left panel (Fig. 1, arrow 1) and click on the table to retrieve a table that provides a quick view of the data available for each organism (Fig. 2a) including the data provider and version.

2.1.1 Genome Sequences and Annotation

1. CryptoDB contains 13 annotated genome sequences including sequences for 9 *Cryptosporidium* species (including several strains), the closely related *Gregarina niphandrodes* [6] and the proto-apicomplexans *Chromera velia* and *Vitrella brassicaformis* [7]. None of the genome sequences is fully complete, telomere-to-telomere without gaps. Contigs have been assigned to scaffolds, super-scaffolds, and chromosomes by the data providers as possible. Currently, *C. hominis* UdeA01, *C. parvum*

Fig. 2 Data summary table and downloads. (a) Use the Data Summary Table, accessed from the Data Summary tab on the left side of the home page, for a quick overview of the data types available for each organism. (b) Genome scale data files are accessed from Downloads, Data Files section in the gray menu bar. Previous releases are archived, and organism folders contain data that cannot be unambiguously assigned to species, such as ESTs

Iowa II, and *C. tyzzeri* UGA55 are represented as 8 chromosomes each, with gaps and at times some left over, unplaced contigs. It is important to bear in mind the relative completeness of each of the sequences when searching the database. The observation that a particular gene is not found in a species does not necessarily indicate that it is not present, as the genome sequences do contain gaps, in some instances hundreds of gaps. The same caution applies to gene or sequence copy number. Paralogous gene families are abundant in the Apicomplexa and parasitic protists in general. Genome assembly algorithms often overassemble multiple copies of highly similar genes into a single gene or are unable to assemble the additional family members if the similarity is lower. In such cases, additional gene family members are often found in the unassembled contigs for the species. This happened with the *Toxoplasma gondii* rhoptry proteins [8].

As there has been very little functional genomic data to assist with genome annotation until very recently, *Cryptosporidium* annotations are also underinformed. The lack of strand-specific RNA-seq or cDNA experimental data has led to under annotation of intron sequences and misannotation of transcripts with overlapping untranslated regions (UTRs). As a user of this and other genomic resources, please bear in mind these rather universal caveats regarding the completeness the genome and annotation.

2. Downloading genome-scale data: The downloads section is accessed from the gray menu bar located near the top of the browser page (see Fig. 1, arrow 2 for the bar and Fig. 2b). This section contains sequence data from the current release as well as archived files from previous releases. Contents are arranged by organism. Folders without a species designation contain data, such as expressed sequence tags, that cannot be assigned unambiguously to a strain. Species folders for unannotated genome sequences contain the genome sequence and predicted open reading frames (ORFs) in fasta format. Genome sequences that are annotated have additional fasta files that contain the predicted protein sequences, transcripts, coding sequence and ORF amino acid translations, as well as GAF- and GFF-formatted annotation files and text files of codon usage, and protein domain associations.
3. Genome analyses/predictions performed at CryptoDB: Sequence and annotation files are integrated into CryptoDB from repositories such as GenBank [9, 10] or, in rare cases, obtained directly from researchers. CryptoDB runs a series of analyses that supplement the existing information and annotation with data such as protein domain prediction, synteny information, and protein structure and function prediction (Table 1).

Table 1
Genome analyses and access in CryptoDB

Analysis	Data	Data category	Appearance
WU-BLAST	Sequence similarities in NRDB (genomic or protein sequence vs. NRDB)	Protein features and properties	Records
HMM-Pfam	Domains	Protein features and properties	Search, records
Protein PI	Isoelectric points for annotated proteins	Protein features and properties	Search, records
Splign	Exon-intron boundaries, splice-junctions, potential frameshifts, and alternative models when there is more than one possibility	Gene models	Search, records,
TMHMM	Predicted transmembrane domain topology	Protein features and properties	Search, records
SignalP	Predicted signal peptides and their likely cleavage sites	Protein features and properties	Search, records
Repeat masker	Interspersed repeats and low complexity DNA sequences	Sequence composition, complexity, and repeats	GBrowse
Tandem repeat finder	Two or more adjacent, approximate copies of a pattern of nucleotides	Sequence composition, complexity, and repeats	GBrowse
Low complexity regions	Sequences containing regions of low complexity are masked with the character “x”	Sequence composition, complexity, and repeats	GBrowse
Pspired	Secondary structure prediction	Structure analysis, protein features, and properties	Search, records
tRNAScan	Predicted tRNA genes	Gene models	Search
InterProScan	Domains and functional site assignments from known proteins	Protein features and properties	Search, records
CAP4 EST Assembly	EST assemblies	Sequence analysis	Records, GBrowse
BLAST EST alignments		Sequence analysis	Records, GBrowse
OrthoMCL	Groups based on clustered ortholog pairs from 150 proteomes across the tree of life	Orthology and synteny	Search, records, GBrowse
ORF generation	Translations in all six reading frames	Sequence analysis	GBrowse

The OrthoMCL [11] analysis creates ortholog profiles across all genome sequences in CryptoDB and beyond. An initial analysis of 150 genome sequences gathered from across the tree of life included several *Cryptosporidium* genome sequences and consisted of an all-vs-all BLASTP of quality screened protein sequences followed by clustering of reciprocal best BLAST hits. The analysis produces groups of similar proteins that include proteins that span the tree of life. The ortholog groups created in this analysis facilitate comparative genomic analysis.

2.1.2 Isolates

1. PopSet Isolates: CryptoDB integrates isolate sequences from the NCBI PopSet database [5]. These sequences originate from phylogenetic, population, environmental, or mutational studies and represent one facet of the known variation for a specific gene, often well-characterized genes commonly used for genotyping like SSU 18S rRNA, *gfp60*, and *comp*.
2. Whole-genome sequencing: Decreasing sequencing costs and improved technologies enable whole-genome sequencing (WGS) on a larger scale, and CryptoDB has integrated several WGS projects as isolates of *C. parvum* and *C. hominis* [12–15]. The isolate data are aligned to the reference genome and analyzed for single-nucleotide polymorphisms (SNPs) and copy number variation (CNV).

2.1.3 Functional and Structural Predictions

1. Gene Ontology assignments: GO terms and their associated identifiers are part of the Gene Ontology controlled vocabulary that describe the biological process, molecular function, and cellular component of a gene product. GO terms in CryptoDB are either manually curated or electronically inferred from orthologs.
2. EC numbers: Enzyme commission numbers are assigned to genes and classify the enzymatic activity of the gene product. CryptoDB receives EC numbers with genome annotation and supplements this with assignments retrieved from other sources [16, 17].
3. InterPro domains: Protein domains and functional insights associated with them are assigned to predicted proteins using InterProScan [18]. The presence of annotated protein domains facilitates sequence- and feature-based searching of identifiable features in known proteins possible. Several protein motif signatures are available in CryptoDB, including Pfam [19], InterPro [20], and Smart [21].

2.1.4 Functional Data

CryptoDB maps functional genomics data (Express Sequence Tag (EST), RNA sequence data, genome-wide RT-PCR, proteomics) [22, 23] to the genome sequence. These data sets document transcription and translational events under different experimental conditions.

1. EST: ESTs represent portions of the sequences expressed by an organism and result from single-pass sequencing of a clone from a cDNA library. ESTs in CryptoDB are downloaded from NCBI's database of ESTs [24] and aligned to genomes.
2. RNA Seq: RNA sequence data represent the full complement of RNA in the sample at the time it was collected, e.g., the intracellular transcriptome. RNA sequence reads are downloaded from archival repositories like the SRA [25] and analyzed through an in-house pipeline that aligns the reads, calculates expression values, and determines differential expression [26–28].
3. Microarray: Spotted glass slide as well as oligonucleotide array data are obtained from the research community and analyzed to determine expression values with an in-house workflow [29].
4. RT-PCR: A large-scale RT-PCR study [30] determined expression values for 3281 *C. parvum* genes.
5. Proteomics: CryptoDB integrates peptide sequence, abundance, and translational modification data from published studies as well [31–34].

2.1.5 Metabolic Pathways and Compounds

Metabolic pathways are downloaded from KEGG [35–37] and MetaCyc [38–40] as networks of enzymes (EC numbers) and compounds (Compound IDs from PubChem Compounds [41]). Genes that are annotated with EC numbers contained in the metabolic pathway are mapped to the pathways, creating a network of gene, pathway, and compound records. Using the strategy system described below, any list of genes can be transformed into the pathways or compounds associated with these genes.

2.2 Home Page and Basic Functions

2.2.1 Layout of Home Page

The CryptoDB home page (Fig. 1) provides quick access to the four data mining resources—record pages, genome-wide searches and strategies, analysis tools, and visualization tools. The page is organized into four sections. The header (Fig. 1a) provides easy access to gene ID (Fig. 1a, arrow 3) and text searches (Fig. 1a, arrow 4) that return genes for further exploration. The header also contains links (Fig. 1a, arrow 5) for registration, login, and a “Contact Us” form for emailing EuPathDB’s active support hotline with questions or suggestions. The gray menu bar (Fig. 1a, arrow 2) provides access to all CryptoDB functionality and remains visible on all CryptoDB pages. Registered users can bookmark record pages in *My Favorites*, save searches and strategies in *My Strategies*, add comments on record pages (Fig. 7), and use the EuPathDB Galaxy site [42] for private analysis of large-scale data accessed through *Analyze My Experiment*. Hover your computer cursor over the *New Search* tab to reveal a dropdown menu for

accessing all searches. The *My Basket* tool provides a mechanism to save records for later viewing or strategy integration. The side panel (Fig. 1b, red headers) contains links to useful information including a summary of data in CryptoDB (Fig. 2a), community resources, educational materials including exercises and video tutorials, news documenting updates implemented with each release, and the EuPathDB twitter feed. The center panels (Fig. 1c, blue headers) contain expandable categories for accessing all searches that return genes (Fig. 1c, arrow 6), searches that return other records (Fig. 1c, arrow 7), and links to useful tools (Fig. 1c, arrow 8) such as BLAST for retrieving records based on sequence similarity, the Companion annotation tool [43] for easy, first-pass, reference-based genome annotation, and the Genome Browser for visualization of sequence-based data aligned to a reference genome. The footer (Fig. 1d) is also available from any page and provides links to the home pages of all EuPathDB sites.

2.2.2 Retrieval of Records Based on Gene ID Searches

Genome sequence annotation specifies a unique identifier for each gene in the genome sequence, and CryptoDB offers several ways to retrieve genes based on their ID. The Gene ID search in the header (Fig. 3a, expansion) navigates directly to the record page of the corresponding gene ID (locus tag). The header appears on every page, making it convenient to reach gene records from any page on the site. A search capable of retrieving multiple records from a list of input gene IDs is available in the *Search for Genes* panel under Annotation, curation, and identifiers (Fig. 3b, arrow 1). Popset Isolate, Genomic sequence, SNP, EST, ORF, Metabolic Pathway, and Compound records can also be retrieved with similar ID searches located in the home page Search for Other Data Types panel (Fig. 3b, e.g., arrows 2, 3).

1. *Visit the gene record page for the gene encoding GP-60.* To visit the record page for the isolate typing gene Glycoprotein _GP40, also known as sporozoite antigen *gp40/15* or *gp60*, enter the gene ID (cgd6_1080) into the *Gene ID* search in the header and then click the search icon to the right of the box. (See Subheading 2.3 for discussion of the gene record page).
2. *Retrieve multiple genes from the database:* Navigate to the *Identify Genes based on Gene ID(s)* search (Fig. 3c) using the home page *Search for Genes* panel (Fig. 3b, arrow 1), enter the following gene IDs (cgd7_230, cgd6_1080, cgd5_4560, GY17_00002758, CTYZ_00003987) in the text box or upload a text file containing the IDs (sample in Electronic Supplemental File 1), and click *Get Answer*. The search results, 5 genes corresponding to the IDs entered in the *Gene ID input set* parameter, appear in the *My Strategies* section (Fig. 3d), which begins with a graphic representation of the search and a summary of results in the top strategy panel (Fig. 3d, blue

Fig. 3 Find records based on IDs. (a) The header Gene ID search navigates directly to the gene page for the ID entered. (b) Multiple records may be retrieved using the dedicated ID searches for each record type, for example,

box). Several options for strategy actions appear at the right of the panel and include an option to share the strategy via a dedicated URL. The URL for this strategy is <http://cryptodb.org/cryptodb/im.do?s=31c90c7080226e9c>. The organism table (Fig. 3d, purple box) shows the distribution of results across the organisms queried. Clicking a number in the organism table filters the result to display only hits from the selected organism. Yellow highlighting in the strategy panel and organism table indicates the active result that appears in the *Gene Results* table (Fig. 3d, brown box). The *Gene Results* section tabulates and displays data for the active result (highlighted with yellow). All CryptoDB data associated with genes, such as GO terms or gene expression values, are available for display in the *Gene Results* tab via the *Add Columns* tool (Fig. 3d, arrow 4). The download tool (Fig. 3d, arrow 5) retrieves either custom (one row per ID) or preconfigured tab-delimited TXT files of chosen data, sequences in FASTA format or annotations in GFF3 format for the active result. The *Genome View* tab (Fig. 3d, red box) displays genomic sequences (contigs or chromosomes) “painted” with the genes from the active result.

2.2.3 Record Retrieval Based on Text

CryptoDB text searches are available for gene, Popset isolate, and compound record retrieval. These searches query the textual content of each record to assemble a set of genes that contain a user-defined text term or phrase. For example, genome annotation, as well as other data in the gene record (e.g., UniProt [17, 44] data, protein domain predictions), contains text that can be used to retrieve genes with similar characteristics.

CryptoDB offers two searches that return genes based on the text in their records: the *Gene Text Search* is available in the header (Fig. 3a, arrow 6) and the dedicated search *Identify Genes based on Text* (*product name, notes, etc.*) is available in the *Search for Genes, Text* (*product name, notes, etc.*) panel (Fig. 3b, arrow 7). A text term or phrase (within quotations) entered in the header *Gene Text Search* initiates a preconfigured search that queries a broad range of data fields in all gene records of all annotated genome sequences in CryptoDB. The dedicated gene text search (Fig. 3b, arrow 7) offers control over the organisms and fields that will be searched. For example, the dedicated gene text search can be configured to search only the annotated gene product name of genes in *C. parvum* Iowa II. Both searches support a wild card (wild card = * and means any character) used for broadening the search to include partial text. For example, a text search of the term *fructo** will

Fig. 3 (continued) genes (arrow 1), Popset IDs (arrow 2), and ESTs (arrow 3). (c) Gene ID search page configured to return 5 genes. (d) My Strategies page depicting results of the ID search shown in (c) and consisting of the strategy panel summary and actions (blue box), the organism filter (purple box) showing the distribution of IDs, and the result table (brown box) with returned IDs and links for downloading and viewing results

retrieve genes whose records contain words that start with the prefix “fructo.” Surrounding a multiword phrase in quotation marks configures the search to find the exact phrase. For example, searching for “*protein kinase*” will return genes with the exact term “protein kinase”; whereas searching for *protein kinase* (without quotations) will return genes with the term “protein,” genes with the term “kinase,” and some may have the phrase “protein kinase.”

1. Use the Gene Text Search to retrieve genes that are likely kinases and download the GeneIDs and download the IDs, product description, and coding sequence length for all results. Navigate to the dedicated gene text search from the home page (Fig. 3b, arrow 6) and enter the term kinase in the box for the Text term parameter (Fig. 4a). Notice the default settings for the Organism parameter (all organisms selected) and Fields (exclude similar proteins (BLAST hits v. NRDB/PDB)). Click Get Answer to initiate a search of all annotated genome sequences for genes whose records contain the term kinase.
2. Examine the search results. The result (over 3500 genes) appears in the My Strategies section and genes from every organism are returned by the search as seen in the Organism table. Results may vary in subsequent database releases due to annotation changes and new database content. The URL for this strategy is <http://cryptodb.org/cryptodb/im.do?s=4be050ac9f4150fe>.
3. Download the product description and coding sequence length for your result. Briefly described above, the download tool is available from any result table. Click Download (Fig. 4b, arrow 1) to open the tool (Fig. 4c). Choose Tab Delimited Excel—Choose columns to make a custom table to retrieve the custom table tool. Since IDs are downloaded by default, choose the Product Description and CDS length as columns, change the Download Type to Text File, and click Get Genes (Fig. 4b, arrow 2). An example of the text file downloaded from CryptoDB 39 is in Electronic Supplemental File 2.

2.2.4 Record Retrieval Based on Sequence Similarity, BLAST Search

Laboratory and bioinformatic research often generate sequence data for genes and isolates. The CryptoDB BLAST search (Fig. 5) employs the NCBI-BLAST algorithm deployed locally on the CryptoDB collection of organisms, to retrieve records based on similarity to a user-defined input query sequence. Input sequences can be compared to CryptoDB genomic sequences, transcripts, proteins, Popset Isolate sequences obtained from GenBank, ESTs, and ORFs. Furthermore, an option exists for comparing an input

Fig. 4 (continued) (*Fields* = exclude Similar proteins (BLAST hits vs. NRDB/PDB)). **(b)** Search results in the My strategies page showing the link to the Download Tool (arrow 1). **(c)** Download tool configured to retrieve a text file containing IDs, product descriptions, and coding sequence lengths for 3519 genes returned by the text search in **(a)**

A Identify Genes based on Text (product name, notes, etc.)

Organism

13 selected, out of 13

Filter list below...

- ☒ Apicomplexa
- ☒ Chromerida

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#)

Text term (use * as wildcard)

kinase

Fields

- ☒ Alias
- ☒ EC descriptions
- ☒ Gene ID
- ☒ Gene notes
- ☒ Gene product
- ☒ Gene name
- ☒ GO terms and definitions
- ☒ Protein domain names and descriptions
- ☒ PubMed
- ☐ Similar proteins (BLAST hits v. NR0B/PDB)
- ☒ User comments

[select all](#) | [clear all](#)

Advanced Parameters

[Get Answer](#)

[Give this search a name \(optional\)](#)

[Give this search a weight \(optional\)](#)

B

My Strategies: [New](#) [Opened \(2\)](#) [All \(71\)](#) [Basket](#) [Public Strategies \(7\)](#) [Help](#)

Hide search strategy panel

(Genes)

[Text](#) [Add Step](#)

Step 1

3519 Genes from Step 1 [Reverse](#)

Strategy: Text(4)

[Click on a number in this table to limit/filter your results](#)

All Results	Ortholog Groups	Apicomplexa									
		C. andersoni	C. hominis	C. megalocytis	C. muris	C. parvum	C. tyzzeri				
		isolate 30847	isolate 30976	isolate TU502_2012	TU502	UdaA01	strain UKMEEL1	RN06	Isos II	isolate UGA05	
3519	047	150	153	169	130	149	151	153	155	157	

Gene Results [Genome View](#) [Analyze Results](#)

Genes: 3519 **Transcripts: 3557** [Show Only One Transcript Per Gene](#)

[First](#) [1](#) [2](#) [3](#) [4](#) [5](#) [Next](#) [Last](#) [Advanced Paging](#) [Download](#) [Add to Basket](#) [Add Columns](#)

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description
Cvrl_15056	Cvrl_15056.11	C. vella CCMP2878	Cvrl_scaffold1097.154.10.425(+)	Ankyrin-3, putative
Cvrl_17561	Cvrl_17561.11	C. vella CCMP2878	Cvrl_scaffold1410.22.860.30.116(-)	Putative ankyrin repeat protein RF_0361, putative

C Download 3519 Genes

Results are from search: Text (product name, notes, etc.)

- Choose a Report:**
- ☒ Tab delimited (Excel) - choose columns to make a custom table
 - ☐ Tab delimited (Excel) - choose a pre-configured table
 - ☐ FASTA (sequence retrieval, configurable)
 - ☐ GFF3: Gene models and optional sequences

Note: IDs will automatically be included in the report and the report will be sorted by ID.

Choose Columns

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#)

Search Columns...

- ☐ Search Specific
- ☐ Gene models
- ☒ Annotation, curation and identifiers
 - ☐ Gene Name or Symbol
 - ☐ Gene has Unmatched Transcripts
 - ☒ Product Description
 - ☐ Transcript Product Description
 - ☐ gene_source_id
- ☐ Link outs
- ☐ Genomic Location
- ☐ Taxonomy
- ☐ Orthology and synteny
- ☐ Genetic variation
- ☐ Transcriptomics
- ☐ Sequences
- ☒ Protein features and properties
 - ☒ CDS Length
 - ☐ Isoelectric Point
 - ☐ Molecular Weight
 - ☐ Protein Length
- ☐ Protein targeting and localization
- ☐ Function prediction

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#)

Choose Rows

- ☐ Include only one transcript per gene (the longest)

Download Type

- ☒ Text File
- ☐ Excel File*
- ☐ Show in Browser

Additional Options

- ☒ Include header row (column names)

[Get Genes](#)

Fig. 4 Dedicated text search page and download tool. (a) Search page shown configured to return genes from all annotated genome sequences (*Organism* parameter = 13 selected out of 13) that contain the word “kinase” (*Text term (use * as wildcard)* = kinase) from the majority of text fields in the gene record Fig. 4

A

Target Data Type

- Transcripts
- Proteins
- Genome
- EST
- ORF
- PopSet **1**

BLAST Program

- blastn
- blastp
- blastx
- tblastn
- tblastx

Target Organism

1 selected, out of 4

Filter list below...

- ☐ Chromerida Popset/Genbank Isolates
- ☐ Cryptosporidiidae Popset/Genbank Isolates
- ☒ Cryptosporidiidae SSU_18S rRNA Reference Isolates
- ☐ Gregarinidae Popset/Genbank Isolates

select all | clear all | expand all | collapse all

Input Sequence

AAGCTCGTAGTTGGATT

Note: only one input sequence allowed.
maximum allowed sequence length is 31K bases.

Expectation value

10

Maximum descriptions/alignments (V=B)

50

Low complexity filter

no

Get Answer

B

My Strategies: [New](#) [Opened \(1\)](#) [All \(1\)](#) [Basket](#) [Public Strategies](#)

Hide search strategy panel

(PopSeqs)

BLAST
50 PopSeqs
Step 1

Add Step

50 Popset Isolate Sequences from Step 1 [Revise](#)

Strategy: **BLAST** **3**

BLAST Popset Isolate Sequences Popset Isolate Sequences Geographical Location

BLASTN 2.2.28+

Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Database: Cryptosporidiidae_SSU_18S rRNA Reference Isolates.fasta
73 sequences; 78,063 total letters

Query= MySeq1
Length=435

		Score	E
AF093490	organism=Cryptosporidium parvum description=Crypto...	785	0.0
AF164102	organism=Cryptosporidium parvum strain IOWA descri...	785	0.0
AF093491	organism=Cryptosporidium hominis renamed from C. pa...	762	0.0
AF112571	organism=Cryptosporidium tyzzeri - renamed from C. p...	760	0.0
AF112572	organism=Cryptosporidium parvum ferret genotype d...	756	0.0

C

BLAST Popset Isolate Sequences Popset Isolate Sequences Geographical Location

First 1 2 3 Next Last Advanced Paging

Download Add to Bask

Popset Sequence Id	Organism	Score	E-Value	Gene Product	Strain
AF093490	Cryptosporidium parvum	785	0E0	small subunit ribosomal RNA	Bovine C. I (BOH6)
AF164102	Cryptosporidium parvum	785	0E0	18S ribosomal RNA	IOWA
AF093491	Cryptosporidium parvum				
AF112571	Cryptosporidium parvum				
AF112572	Cryptosporidium parvum				
AF115378	Cryptosporidium wrairi				
AF112574	Cryptosporidium meleagridis				
EF641022	Cryptosporidium sp. beaver				
EF112574	Cryptosporidium sp. 160A				

Map Satellite

Fig. 5 Retrieve records based on BLAST similarity. (a) BLAST search page configured to return Popset records (arrow 1) from a comparison of the input sequence against the *Cryptosporidiidae* SSU_18S rRNA Reference Isolates (arrow 2) database in CryptoDB. (b) Blast search result page in the My Strategies

18S sequence to the community curated *Cryptosporidiidae* SSU_18S rRNA Reference Isolate set to determine the putative identity of the input sequence relative to reference sequences that are considered standards by the *Cryptosporidium* community.

1. *Find the genotype of an unclassified Cryptosporidium isolate.*

Consider that the following sequence originated from a laboratory study of the 18S small subunit ribosomal RNA gene from an isolate retrieved during a *Cryptosporidium* outbreak at a swimming pool. The sequence (below and Electronic Supplemental File 3) was identical from all isolates. Use the CryptoDB BLAST search to find a similar reference sequence.

```
AAGCTCGTAGTTGGATTTCTGTTAATAA
TTTATATAAAATATTTTGATGAATATTTATATAATA
TTAACATAATTCATATTACTATATATTTTAG
TATATGAAATTTTACTTTGAGAAAATTAGAG
TGCCTTAAGCAGGCATATGCT
CTTGAATACTCCAGCATGGAATAATA
TTAAAGATTTTATCTTTCTTATTGGTTCT
AAGATAAGAATAATGATTAATAGGGACA
GTTGGGGGCATTTGTATTTAACAGTCAGAGGTGAA
ATTCTTAGATTTGTTAAAGACAACTAATGCGAAAG
CATTTGCCAAGGATGTTTTTCATTAATCAAGAACG
AAAGTTAGGGGATCGAAGACGATCAGATACCGT
CGTAGTCTTAACCATAAACTATGCCAACT
AGAGATTGGAGGTGTTCTCTTA
TCCTTCAGCACCTTA
```

2. Navigate to the BLAST search from the tools menu and set the *Target Data Type* to Popset (Fig. 5a, arrow 1) which will configure the query to return Popset Isolate records and to use the BLASTN algorithm. For *Target Organism*, choose *Cryptosporidiidae* SSU_18S rRNA Reference Isolates (Fig. 5a, arrow 2) to BLAST your input sequence against the isolate reference database. Paste the sequence into *Input Sequence* and click *Get Answer* to initiate the BLAST.
3. The results appear in *My Strategies* (Fig. 5b) with a graphic summary of the strategy at the top followed by three tabbed pages for the results: BLAST (Fig. 5b, arrow 3), Popset Isolate Sequences (Fig. 5c), and Popset Isolate Sequences Geographic Location (Fig. 5c, inset). The BLAST result (BLAST tab) reveals two hits with 100% identity, to *C. parvum* over the length of the sequence provided. One hit was to a *C. parvum*

←
Fig. 5 (continued) section tabulates the 50 best Popset sequences based on Score and E value. (c) Other tabbed pages in the BLAST search results. *Popset Isolate Sequences* tab lists the record IDs with associated CryptoDB data and offers a sequence alignment tool, while the *Popset Isolate Sequences Geographical Location* displays the collection site location

bovine strain (AF093490) and the other to *C. parvum* Iowa II (AF164102), indicating that the outbreak is likely caused by *C. parvum*.

2.3 Record Pages

CryptoDB's extensive record system is a rich data mining resource. Records compile all database information into tables and graphs for genes and nongene entities including genomic sequences, Popset Isolate Sequences, Genomic Sequences, Genomic Segments, SNPs, ESTs, ORFs, Metabolic Pathways, and Compounds. All record pages open with a summary section at the top (Fig. 6a) and employ a floating, collapsible *Contents* panel (Fig. 6b) on the left for navigating the data section (Fig. 6c) on the right. This section of the chapter will explore the structure of the record page and highlight important data content using record pages for the gene that encodes GP60 (cgd6_1080), the isolate record page for *Cryptosporidium parvum* isolate 830 surface glycoprotein 900 gene, and the metabolic pathway for *Glycolysis/Gluconeogenesis* (KEGG).

2.3.1 Gene Record Page Organization and Content

1. *Use the Gene ID search in the header to open the record page* for the gene encoding GP60 (cgd6_1080) (Fig. 6a, inset). The gene page opens with summary information at the top (Fig. 6a) including gene ID, gene product name, genomic location, and species/strain information. In this case, the official gene name, Glycoprotein_GP40 differs from the gene and protein name commonly used by the research community. See **Step 2** below for more information about gene aliases and other names. Links to bookmark (*Add to favorites*) or save the gene for later use (*Add to basket*) as well as to download sequences or gene data appear at the top of the page (Fig. 6a, arrow 1). The *Shortcuts* (Fig. 6a arrow 2) provide quick visualizations of the gene page data via the magnifying glass icon and navigate to the corresponding image in the data when the image or the label above is clicked. The collapsible *Contents* section (Fig. 6b) directs the data section (Fig. 6c) of the page to the desired content. Uncheck a box to hide that data from the page. Data (Fig. 6, arrow 3) are represented in images, graphs, and downloadable tables (Fig. 6, arrow 4). The *Data sets* link next to the section title reference the data sets from which the data were taken (Fig. 6, arrow 4). *View in genome browser* links offer easy access to the genome browser in the genomic region containing the gene (Fig. 6d). All mapped sequence-based data are available for interactive display via the *Select Tracks* page (Fig. 6d, arrow 5).
2. *Annotation, curation, and identifiers section. Use the Contents navigation tool to view this section* (Fig. 7a), which offers alternative gene product descriptions, aliases, notes from the annotator, and a *User Comments* table. As new versions of genome

Fig. 6 Gene record page for cgd6_1080, Glycoprotein_GP40. (a) Summary section containing basic location and identity information. Quick links to send the gene to Basket and Favorites or download (arrow 1) are available. Shortcuts (arrow 2) direct the data section to display the associated data. (b) Searchable Contents panel for navigating the data on the page showing the data section directed to the Gene Models section. (c) Data section focused on Gene Models which are represented in images (arrow 3) or downloadable tables (arrow 4). (d) Browser tab of the genome browser accessed from the *View in genome browser* button in the gene page data section. The *Select Tracks* tab (arrow 5) is for choosing tracks to display in the Browser tab

A

Contents

expand all | collapse all

Search section names...

- 1 Gene models
- 2 Annotation, curation and identifiers**
- 3 Link outs
- 4 Genomic Location
- 5 Literature

2 Annotation, curation and identifiers

▼ Alternate Product Descriptions [Data sets](#)

No data available

External DB Version 2007-02-26

▼ Names, Previous Identifiers, and Aliases [Download](#) [Data sets](#)

Name/ID/Alias	Type
CPIW_00001274	alternate ID

▼ Notes from Annotator [Data sets](#)

No data available

Product Description Glycoprotein_GP40

▼ User Comments [Download](#) [Data sets](#)

Add a comment

Search this table... Showing 2 rows

Comment ID	Headline	PubMed ID(s)
64240	also known as gp40/15	10858228,17719100,17043102
27850	Also known as GP60	

B

Add a comment to gene cgd6_1080

Please add only scientific comments to be displayed on the gene page for cgd6_1080. If you want to report a problem, use the support page. Your comments are appreciated. They will be forwarded to the genome curators. If this is a **new gene**, please also add a comment in the corresponding

Part I: Comment

Headline *

Category (check all that apply) ☐ Gene Model ☐ Name/Product ☐ Function ☐ Expression ☐ Sequence ☐ Phenotype

Comment *

Location

Strand: ☐ Forward ☒ Reverse

Genome Coordinates: 266434-267560 [help](#)

Part II: Evidence for This Comment (Optional)

Upload File

Select a file: [Choose File](#) No file chosen

Brief Description: (4000 max characters)

Add Another File

PubMed ID(s)

[Preview](#) the article details of the PubMed ID(s) above [help](#)

Search for PubMed IDs <enter valid gene ID>

Digital Object Identifier (DOI) Name(s) [help](#)

Genbank Accession(s) [help](#)

Part III: Other Genes to which you want to apply this comment (Optional)

Gene Identifiers [help](#)

Add Comment

Fig. 7 Gene page section for annotation, curation, and identifiers. (a) Contents panel focused on the Annotation, curation, and identifiers section of the gene page. Previous names are listed in a table (arrow 1) as are comments entered by users that serve to enhance the official genome annotation. (b) Comment form accessed from the “Add a comment” link (long downward arrow) which is used to add information to the gene page and can contain files and images

annotation are accepted by the community and integrated into CryptoDB, gene IDs and product descriptions can change. CryptoDB maps the aliases within table *Names, Previous Identifiers, and Aliases* (Fig. 7a, arrow 1) to the current ID so that an ID search performed with an alias (CPW_00001274) will return the record page of the current ID, cgd6_1080.

CryptoDB encourages the community to enhance the gene page with information added via the *User Comments* platform. Comments can offer valuable additions to the official annotation such as alternative gene models, unpublished data, and associated publications or references. The comment form (Fig. 7b) used for entering comments is retrieved from the *Add a comment* link (Fig. 7a, down arrow). Users must be registered and logged in to enter user comments. Once added, the comment immediately becomes part of the record page and the CryptoDB searchable text. For example, the comments entered here (Fig. 7a, arrow 2) provide the alternative gene names “gp40/15” and “GP60.” Although “GP60” does not appear in the gene product name, it does appear in the user comment which is queried during the *Gene Text Search*, so searching for the text term “GP60” will return the gene: cgd6_1080, Glycoprotein_GP40.

3. *Orthology and synteny section. Explore orthology and synteny for cgd6_1080, Glycoprotein_GP40 and compare the results to the cgd2_60, ABC_transporter.* Navigate to the Ontology and Synteny section with the *Contents* navigation tool and explore orthology for cgd6_1080, Glycoprotein_GP40 (Fig. 8a). The ID of the ortholog group to which this gene belongs appears as a link to the OrthoMCL group record (Fig. 8a). Use this link to go to OrthoMCL and explore the group features (scroll down in OrthoMCL.org to the *List of Sequences* table). Only two sequences are assigned to this group, supporting the species-specific and single-copy nature of cgd6_1080, Glycoprotein_GP40. Out of the 150 proteomes contained in OrthoMCLDB, which included *C. hominis* and *C. parvum* representing the *Cryptosporidiidae*, only one sequence each from *C. hominis* and *C. parvum* is represented in the group.

Visit the Orthologs and Paralogs within CryptoDB table which lists all orthologs from organisms within EuPathDB (Fig. 8b). This table expands the local list of orthologs compared to the Ortholog group at OrthoMCL. As CryptoDB integrates newly available annotated genome sequences, orthology is determined by assigning their proteomes to existing Ortholog groups without updating the original 150 proteome analysis that forms the underlying data set for OrthoMCLDB. Orthologs of cgd6_1080 do not appear in *Gregarina*, *Chromera*, or *Vitrella*. Only one gene from each *Cryptosporidium* species is orthologous and all are syntenic, in the same genomic location,

A

Contents

expand all | collapse all

Search section names...

1 Gene models

2 Annotation, curation and identifiers

3 Link outs

4 Genomic Location

5 Literature

6 Taxonomy

7 Orthology and synteny

8 Genetic variation

7 Orthology and synteny

Ortholog Group OG5_222676

Orthologs and Paralogs within CryptoDB

Retrieve multiple sequence alignment or multi-FASTA

Synteny

OrthoMCL DB

Ortholog Groups of Protein Sequences

Home New Search My Strategies My Basket (0)

Group: OG5_222676

(2 sequences)

Add to Basket Add to Favorites

B

Search this table... Showing 7 rows

Gene	Organism	Product	is syntenic	has comments
Chro.60138	Cryptosporidium hominis TU502	sporozoite antigen gp40/15	yes	no
ChTU502y2012_406g0275	Cryptosporidium hominis isolate TU502_2012	glycoprotein	yes	no
CHUDEA6_1080	Cryptosporidium hominis UdeA01	unspecified product	yes	no
CmeUKMEL1_04910	Cryptosporidium meleagridis strain UKMEL1	Glycoprotein GP40 of Cryptosporidium family protein	yes	no
CTYZ_00001283	Cryptosporidium tyzzeri isolate UGA55	Glycoprotein GP40	yes	no
cubi_02234	Cryptosporidium ubiquitum isolate 39726	hypothetical protein	yes	no
GY17_00000107	Cryptosporidium hominis isolate 30976	Glycoprotein GP40	yes	no

Fig. 8 Orthology and synteny. (a) Navigation route from the gene page Contents tool to the *Orthology and synteny* section of the gene page and finally to the gene's OrthoMCL group page. (b) Orthologs and Paralogs within EuPathDB table for *cgd6_1080*

reflecting the conserved nature of this gene. Also notice the product name does not always reflect the activity of the gene, reflecting the maturity of the annotation. Use the *Gene Text* search in the header to search for *gp40*. The search returns less than 10 genes, one for each organism, again confirming the conserved single-copy nature of the gene.

Contrast the orthology information for *cgd6_1080*, *Glycoprotein_GP40* with *cgd2_60*, *ABC_transporter*. Navigate to the gene record page for *ABC_transporter* by entering the ID (*cgd2_60*) in the *Gene ID* search in the header. Use the

Contents section to view the *Orthology and synteny* data section and notice the *Orthologs and Paralogs within EuPathDB* table. The table contains multiple entries for each *Cryptosporidium* species suggesting that gene duplication events may have occurred.

Return to the *cgd6_1080, Glycoprotein_GP40* gene page by clicking the browser back button or using the Gene ID search in the header. The *Retrieve multiple sequence alignment or multi-FASTA* tool part of the *Orthology and Synteny* section can be used to produce a ClustalW multiple sequence alignment in the region of the gene. Choose all organisms (Fig. 9a, arrow 1) and click *Run alignment*. The alignment appears in a new tab and is an alignment between 9 sequences, the gene itself, and 8 additional genomic sequences found to have similarity. It is interesting to note that the *C. andersoni* sequence in the alignment is missing from the Ortholog table. A closer look at the *Synteny* graph (Fig. 9b), which shows the chromosomal organization of genes across organisms with orthology indicated by shadows connecting the genes, shows that no gene is annotated for the region in *C. andersoni* or *C. muris* (Fig. 9b, red box). Since the MSA returns an alignment for *C. andersoni*, this suggests that a gene should be called for the region and this information would make a good comment for the genomic sequence record page. Also notice that *Gregarina*, *Chromera*, and *Vitrella* do not have syntenic sequences or orthologs to GP60, indicative of the distant evolutionary relationship between the organisms.

4. *Genetic Variation section: Compare variation in the gp60 gene among isolates from China, Egypt, and Uganda.* From the *Synteny* graph, scroll down to the *Genetic Variation* section which contains the *Isolate Alignments in this Gene Region* tool and a summary of *DNA Polymorphism* determined from an analysis of whole-genome sequencing of isolates. The isolate alignment tool (Fig. 10a) produces an alignment of the whole-genome sequence data within the region of the gene. The tool opens with all isolates selected for the alignment and includes the option to reduce the number of isolates included in the alignment based on isolate characteristics such as geographic location of isolate collection or gp60 genotype. The left panel of the tool shows the isolate characteristics, while the right panel shows the data distribution for the characteristic chosen on the left. Navigate the filter tool (Fig. 10a, arrow 1) to choose only isolates from China, Egypt, and Uganda and retrieve the alignment with the *View Results* button. The results appear in a new tab with variation highlighted in pink. Figure 10b shows an area of the alignment with a SNP in the Ugandan isolate (TU114) (Fig. 10a, arrow 2) and another area where all but one Egyptian isolate (35090) diverge from

A

Contig ID:

Nucleotide positions: to ☐ Reverse & complement

Organisms to align:

Select 15 or fewer organisms from the tree below.
You have currently selected 13

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#)

Search for Organism(s) to include in the alignment or expand the tree below

☒ Apicomplexa 1

☒ Chromerida

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#)

Select output:

☒ Multiple sequence alignment (clustal)

☐ Multi-FASTA

[Run alignment](#)

CLUSTAL O(1.2.3) Multiple Sequence Alignments

```
(LR501000008) 128153 TTACACCGC GGTGAAGCTG ACAGATTTCG GATATGGGAA
(LR501000017) 650957 TTAAACATG AAGATTCACG CGAGATTTCG AAATACGAC
218K01000008 87989 .TACATCATG AATAAGCTG CGAGATTTCG AAAACCGAA
LN877952 259332 .TACACACG AATAAGCTG CAAGATTTCG AAAACCGAA
(218K01000101) 362766 TTACACACG AATAAGCTG CAAGATTTCG AAAACCGAA
(218K01000007) 3816125 TTACACACG AATAAGCTG CAAGATTTCG AAAACCGAA
(AEL01000025) 8815 TTACACACG AATAAGCTG CAAGATTTCG AAAACCGAA
CM000434 266434 TTACACACG AATAAGCTG CAAGATTTCG AAAACCGAA
cty2_6 254539 .TACACACG AATAAGCTG CAAGATTTCG AAAACCGAA

Cvel_scaffold572 3785 .....
(LR501000008) 128173 CTTTCAATTT AACTGTGTTG TTATGTTCTC CATAGTCTTA
(LR501000017) 650887 CTTTCAAAAG GACGATATTG GAAGATTCAC CTAGTTCACC
218K01000008 88068 CTTTCAAAAG AACTGTGTTG TCCTTGTCCA CCAAGTTCTC
LN877952 259411 CTTTCAAAAG AACTGTGTTG TTCTTGTCCA CCAAGTTCTC
(218K01000101) 362686 CTTTCAAAAG AACTGTGTTG TTCTTGTCCA CCAAGTTCTC
(218K01000007) 3816445 CTTTCAAGAG AACTTTGTTG TCCTTGTCCA CCAAGTTCTC
(AEL01000025) 8735 CTTTCAAGAG AACTTTGTTG TCCTTGTCCA CCAAGTTCTC
CM000434 266514 CTTTCAAAAG AACTTTGTTG TCCTTGTCCA CCAAGTTCTC
cty2_6 254618 CTTTCAAAAG AACTTTGTTG TCCTTGTCCA CCAAGTTCTC

Cvel_scaffold572 3786 AGACACGAGA AAGGTTCAAG GC.....
(LR501000008) 128093 ATTTGCGCAG AATGTAAAGG TTAATTTTCT GGTACAAATG
(LR501000017) 650907 TTATGTCGAG TATAGATATG CTGGTTATCA GCGATTAAC
218K01000008 88148 TTGCGCCGAG TATAAAGT TTATGTTCTC GCAACCAAA
LN877952 259491 TTGCGCCGAG TATAGAAAGT CTTATGCTCT GCAACCAAA
(218K01000101) 362486 TTGCGCCGAG TATAGAAAGT CTTATGCTCT GCAACCAAA
(218K01000007) 3816165 TTGCGCCGAG TATAGAAAGT CTTATGCTCT GCAACCAAA
(AEL01000025) 8655 TTGCGCCGAG TATAGAAAGT CTTATGCTCT GCAACCAAA
CM000434 266594 TTGCGCCGAG TATAGAAAGT CTTATGCTCT GCAACCAAA
cty2_6 254698 TTGCGCCGAG TATAGAAAGT CTTATGCTCT GCAACCAAA
```

Fig. 9 Gene alignments and synteny on the gene page. (a) Use the *Retrieve multiple sequence alignment* or *multi-FASTA* tool in the “Orthology and synteny” section for quick alignments in chosen organisms (arrow 1). (b) Synteny graph on the gene page is a static representation. Click the “View in genome browser” button to navigate to GBrowse with the possibility to change settings for the synteny graph

Fig. 10 Isolate alignments and SNPs on the gene page. (a) *Isolate Alignments in this Gene Region* tool for retrieving alignments of isolate whole-genome sequence data with options for choosing isolates based on isolate characteristics (metadata). As shown, the tool is configured to align isolates from China, Egypt, and

the reference *C. parvum* sequence (CM000434) (Fig. 10a, arrow 3).

Close the Isolate Alignment tool by clicking the triangle next to the title (Fig. 10a, arrow 4) and notice the *DNA polymorphism* section which includes a summary of the coding potential, the incidence of all SNPs across all isolates, and an image of SNPs mapped to the gene model. In your browser, hover your mouse over the SNPs in the image to reveal a panel of detailed information including a link to the SNP record page (Fig. 10c).

5. *Transcriptomics Section: Scroll down from the DNA polymorphism section to view the Transcriptomics table* (Fig. 11), which includes all transcript expression data sets with data that map to this gene. Data types include genome-wide RT-PCR, microarray, and RNA-seq data. Table rows expand to reveal the results of CryptoDB transcriptomic analysis presented in graphs and tables. Click the triangle in the header next to the data set “Expression profiling of life cycle stages post-infection” to view the data graphs and tables (Fig. 11, arrow 1). For this experiment, transcript abundance was determined using Real Time-PCR for >3200 genes over a 72-h infection of HCT8 cells [30]. The data graph reveals that the gene is maximally expressed at 12 h postinfection.

Compare the oocyst and intracellular stage expression for this gene. A comparison of oocyst and intracellular stage transcriptomes is also available, and the graphs and tables associated with this RNA-sequence analysis can be viewed by expanding “Transcriptome of oocyst and intracellular stages” (Fig. 11, arrow 2). The data indicate that the gene is highly expressed in the intracellular stages with very little expression in oocysts. Open *Coverage* (Fig. 11, arrow 3) to reveal a snapshot of the RNA sequence coverage plots and these can be further interrogated using the *View in genome browser* link.

6. *Proteomics section: Use the Content section to navigate to the proteomics data.* Proteomics data are tabulated in the *Mass Spec.-based Expression Evidence* table as well as associated data such as sequence counts and observed spectra. The *Mass Spec.-based Expression Evidence Graphic* depicts individual peptides mapped to the protein sequence. In your browser, using your mouse to hover over the rectangle peptide glyphs retrieves details concerning the mapped peptide sequence and its spectral count and the experiment in which it was detected (Fig. 12).

Fig. 10 (continued) Uganda. Isolates are chosen based on sample metadata, in this case the country of isolate collection (arrow 1) with China, Egypt, and Uganda as values. **(b)** A portion of the resulting alignment indicating variation highlighted in pink. **(c)** SNPs image from the *DNA Polymorphism* section of the gene page. SNPs are colored by coding potential and reveal detailed information upon hover

A

▼ Transcript Expression

Data sets

Search this table... Showing 5 rows

Preview	Name	Summary	Attribution	Assay Type
Expression profiling of life cycle stages post-infection	Expression profiling of <i>Cryptosporidium parvum</i> during in vitro development determined by high throughput RT-PCR.		RT-PCR
<div><div>Median Expr (standardized) - cg6_1080</div> <div><div>Data table</div><div>Description</div><div>X-axis</div><div>Time point</div><div>Y-axis</div><div>Median Expr. (standardized)</div><div>Choose gene for which to display graph</div><div>cg6_1080</div><div>Choose graph(s) to display</div><div>expm_val</div><div>Graph options</div><div>Show log Scale (not applicable for log(ratio) graphs, percentile graphs or data tables)</div></div></div>				
Full Dataset Description
Oocyst Transcriptome: Control vs. UV treatment	Global gene expression in <i>C. parvum</i> oocysts (environmental stage) and oocysts treated with UV were determined by whole genome microarray.	Zhang et al.	array
Transcriptome of sporozoites and HTC-8 infection time course	Transcriptomes of freshly excysted <i>C. parvum</i> sporozoites as well as parasites collected 48- and 96- hours post-infection in HTC8 cells were determined by SOLiD, paired end, strand-specific RNA sequencing.	Lippuner et al.	RNA-seq
Transcriptome of oocyst and intracellular stages	<i>Cryptosporidium parvum</i> transcriptome from intracellular (24 h post-infection) and oocysts Stages.	Widmer et al.	RNA-seq

fpkm_sense - cg6_1080

Data table

Description

X-axis

Intracellular culture 24hrs; Oocysts

Y-axis

Transcript levels of fragments per kilobase of exon model per million mapped reads (FPKM). The percentile graph shows the ranking of expression for this gene compared to all others in this experiment.

Choose gene for which to display graph

cg6_1080

Choose graph(s) to display

fpkm_sense fpkm_antisense percentile_sense percentile_antisense

Graph options

Show log Scale (not applicable for log(ratio) graphs, percentile graphs or data tables)

Full Dataset Description

▼ Coverage

Non-unique mapping may be examined in the genome browser (tutorial)

View in genome browser

Transcriptome of 48 hours in vitro infection

Transcriptome (RNA sequence) from *Cryptosporidium parvum* Iowa isolated from an in vitro infection of cultured HCT-8 cells.

Isaza et al.

RNA-seq

Fig. 11 Transcriptomics section of the gene page. Transcriptomics table with expandable rows (arrows 1 and 2) that reveal detailed information concerning the experimental data. Data are represented in graphs and tables as well as coverage plots for RNA sequence data (arrow 3)

16 Proteomics

► Mass Spec.-based Expression Evidence [Data sets](#)

▼ Mass Spec.-based Expression Evidence Graphic [Data sets](#)

Fig. 12 Proteomics section of the gene page. *Mass Spec.-based Expression Evidence Graphic* expandable table displaying graphic of peptides mapped to the protein sequence for five different proteomics experiments in CryptoDB. Mapping details are revealed when hovering over the peptides glyphs

2.3.2 Isolate Record
Organization and Content

1. Retrieve the record for isolate AF527844, *Cryptosporidium parvum* isolate 830 surface glycoprotein 900 gene, partial cds by entering the ID into the ID search that returns Popset isolate records on the CryptoDB home page (Fig. 13a). Since the

search was run on only one ID, the result navigates to the record page instead of My Strategies. The layout of the isolate record page is similar to the gene page with a summary section, a floating Contents section (which can be closed using the double arrowhead on the right top), and a data section. The summary section contains pertinent information such as isolation source, location, organism, and product description information (Fig. 13b). Links to bookmark, save, and download the record appear above the ID and product description. The sequence section provides the nucleic acid sequence associated with the isolate, while the Blast Similarity Alignments and Overlapping Genes table shows the results of a BLAST of the isolate sequence against all genome sequences in CryptoDB (Fig. 13c). Overlapping genomic sequences and genes are indicated for each BLAST hit.

2.3.3 Metabolic Pathway Record Organization and Content

The Cryptosporidiae are known to be missing several metabolic pathways [45, 46]. While basic pathways such as glycolysis are present, surprisingly, both purine and pyrimidine biosynthesis pathways are missing in Cryptosporidiae, although salvage pathways are present [47]. Amino acid biosynthetic pathways are also missing for lysine, tyrosine, alanine, serine, selenocysteine, threonine, and thiamine. [48] Another pathway, the tricarboxylic acid (TCA) cycle is present in *C. muris* but absent from *C. parvum* and *C. hominis* [48]. These qualities can be investigated in the CryptoDB representation of metabolic pathways.

1. Retrieve the Glycolysis/Gluconeogenesis (KEGG) and explore the basic functions of the page. Navigate to the Pathway Name/ID search under Metabolic Pathways in the Search for Other Data Types category on the home page (Fig. 14a). Set the Pathway Source to KEGG and begin typing “glycolysis” in the Pathway Name or ID parameter and then choose the pathway from the list. Click Get Answer to retrieve the pathway.

The organization of the page is similar to other record pages with a Summary at the top, a collapsible Contents section, and the data presented in images and searchable tables. An interactive Cytoscape [49] image (Fig. 14b) shows the pathway as a series of enzymatic reactions with square enzyme nodes and circular compounds. The image can be repositioned by dragging or with the navigation tool (Fig. 14b, arrow 1). Resizing the image is accomplished with the zoom tool (Fig. 14b, arrow 2) or by scrolling a mouse in or out while hovering over any portion of the image. At low zoom levels, pathway nodes appear as glyphs. At mid zoom levels, the main compound nodes appear as molecular formulas. At high zoom levels, enzymes and side nodes appear as formulas, terms, or Enzyme Commission numbers. Clicking on a node retrieves details

Fig. 13 Popset isolate records. (a) Popset Isolate ID search is accessed from the *Other Data Types* panel, *Popset Isolate Sequences*, *Popset ID(s)* search. Shown configured to return one record—AF527844. (b) Summary section of the Popset Isolate record. (c) *Blast Similarity Alignments and Overlapping Genes* table showing the results of a CryptoDB BLAST analysis of the sequence against all genome sequences integrated into CryptoDB. Links to matching genomic sequences and overlapping genes are provided

Fig. 14 Metabolic pathway record page. (a) Metabolic Pathway Name/ID search is accessed from the *Other Data Types* panel, *Metabolic Pathways*, *Pathway Name/ID*. The search is shown configured to query for KEGG Pathways and use the type-ahead function of the *Pathway Name or ID* parameter.

about the node, regardless of zoom level (Fig. 14b, arrow 3 and expansion). Enzyme nodes highlighted in orange represent enzymes for which at least one gene in CryptoDB is annotated with the EC number of the enzyme. Pathway reactions are tabulated in Metabolic Pathway Reactions (Fig. 14, arrow 4) below the pathway image, and Pathways with Shared Reactions offer links to related pathways.

2. Annotate the Glycolysis/Gluconeogenesis (KEGG) pathway using the Paint Enzymes tool. Click on the yellow part of the glyph representing Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) (Fig. 14b, arrow 3) to retrieve the node details. Open the Paint Enzymes tool in the Cytoscape Drawing menu bar and choose By Genera (Fig. 15, arrow 1). Use the Genera Selector tool to choose *Cryptosporidium*, *Gregarina*, *Chromera*, and *Vitrella*. Click Paint to retrieve a graph depicting which organisms have genes with this EC number. The graph (Fig. 15, arrow 2) indicates that all organisms have at least one gene with this enzymatic activity. To retrieve a list of the genes, click the Show genes which match this EC Number link (Fig. 15, arrow 3) to initiate a search for genes (“Identify Genes based on EC Number”) that are annotated with the EC number. The search returns 25 genes. Scroll to the organism table in the search result and notice that the search returned one gene for each *Cryptosporidium* organism in CryptoDB, while multiple genes annotated with the EC number are found in *Gregarina*, *Chromera*, and *Vitrella*. These are likely paralogs/orthologs since all genes belong to the same ortholog group.

Pathways may also be annotated with expression data from Microarray and RNA sequence data to give a quick overview of pathway expression. Using a similar workflow as above, explore the expression of enzymes within the glycolysis pathway (Fig. 15, inset). Paint the pathway with the RNA sequence data set called “*C. parvum* Iowa II Transcriptome of oocyst and intracellular stages (Widmer et al.)” in which the authors determined the transcriptomes of oocysts and *C. parvum*-infected cultures.

3. Explore the TCA cycle for *Cryptosporidium* species. Navigate to the MetaCyc pathway record for TCA cycle I (prokaryotic) as above and annotate the pathway by genera with *Cryptosporidium*, *Gregarina*, *Chromera*, and *Vitrella*. Click on the nodes to inspect the genera graphs and notice that at least

Fig. 14 (continued) **(b)** Glycolysis/Gluconeogenesis pathway record page depicting the Cytoscape representation. Navigation (arrow 1) and zoom (arrow 2) tools are available and will reveal enzyme EC numbers and compound structures at sufficient zoom. Node details appear once an enzyme (square) or compound (circle) has been clicked. *Metabolic Pathway Reactions* (arrow 4) tabulate all reactions in the pathway

Fig. 15 Annotate metabolic pathways. (a) Metabolic pathways can be annotated with orthology and functional genomics data using the *Paint by Genera* (arrow 1) and *Paint by Experiment* tools, respectively. Once annotated, *Node Details* will include data graphs (arrow 2 and inset). The *Show gene(s) which match this EC Number* (arrow 3) link initiates a search for genes with that EC number

one gene encoding each enzyme is present in *Cryptosporidium*. To learn the distribution of enzymes across the CryptoDB organisms, use the Show genes with matching EC numbers link within the node details panel or the link in the Gene Count column of the Metabolic Pathway Reactions table. The organism table in the search result page reveals that *C. muris* and *C. andersoni* have genes representing the enzyme EC 1.1.1.42, but it is missing from the other *Cryptosporidium* strains. Using the table links to investigate the remaining enzymes in the pathway, *C. muris* is the only *Cryptosporidium* species to have all the enzymes, supporting the finding that the TCA cycle exists in *C. muris* but is missing from *C. parvum* and *C. hominis* [48, 50].

2.4 Searches, Strategies, and Result Analyses

The CryptoDB strategy system is a unique and powerful tool for exploring relationships across data sets, data types, and organisms. The system offers over 100 preconfigured searches that query individual data sets ranging from genome-wide analyses for predicted signal peptides to GO annotation, transcriptomic analyses, or SNP analyses. Each search provides evidence for a specific biological property, returning a list of records that meet the search criteria and therefore share the biological property intrinsic to the data set. For a comparative genomics perspective, searches can be configured to query any or all genome sequences integrated into CryptoDB. The system can easily answer questions concerning stage-specific expression, expression timing, biological function, and more.

Strategies are created by adding, subtracting, intersecting, joining, or collocating the results of individual searches. An orthology transform tool can convert a gene result in one organism to their orthologs in another, making it possible to take advantage of a data-rich organism such as *C. parvum* to garner information about less studied organisms such as *Gregarina*. A colocation tool is used to explore relationships based on relative genomic location, for example, to find SNPs located within 500 bp of annotated genes. A nesting tool can be used to control the strategy logic while combining search results. The “Analyze Results” tools offer GO, metabolic pathway, and gene product word enrichment analyses for any gene search or strategy result. Isolate search results can be analyzed by MSA. The following examples illustrate how to create and analyze a gene strategy and an isolate strategy in CryptoDB.

2.4.1 Gene Strategy Example

This example assembles a list of *G. niphandrodes* genes that are likely kinases expressed in oocysts by creating a four-step strategy. The first step is a text search that retrieves genes from all organisms in CryptoDB whose records contain the word kinase (Fig. 16a, arrow 1). The second search (Fig. 16a, arrow 2) retrieves genes annotated with GO terms associated with kinase activity which provides a second line of evidence for kinase activity. The first two

Fig. 16 Example 1 strategy, Step 1 result and adding Step 2. (a) Strategy panel depicting the final four-step strategy with search (arrows 1, 2, 4) and step results (arrows 3, 5, 6). (b) Strategy result after the text search for kinase and the *Add Step* panel for extending the strategy. (c) GO Term search page accessed from the *Add Step* panel configured to combine the GO Term search with previous results using a union operator. (d) The two-step strategy results after running the GO Term search and combining it with the text search

searches are combined by adding (union) the two search results to create the Step 2 result (Fig. 16a, arrow 3), which is a more comprehensive set of kinases across all organisms. Next, we want to restrict the list of kinases to only those with evidence of expression during the oocyst stage. Since no transcriptomic data are publicly available for *G. niphandrodes*, we will query *C. parvum* data and then transform those results into *G. niphandrodes*, thus inferring expression is conserved between organisms (potentially an invalid assumption, but it serves for illustrative purposes). The third search in the strategy (Fig. 16a, arrow 4) queries the expression values determined in an RNA sequence analysis comparing the oocyst and intracellular stages of *C. parvum*. The oocyst expressed genes are intersected with the kinases to produce a Step 3 result (Fig. 16a, arrow 5), including kinases that are also expressed in the oocyst stage. An intersection returns genes in common between the two search results. Since Step 2 is a multi-organism set of kinases, while Step 3 contains only *C. parvum* oocyst genes; the Step 3 result (intersection) contains only *C. parvum* genes. The fourth step (Fig. 16a, arrow 6) transforms the *C. parvum* oocysts-expressed kinases into their *G. niphandrodes* orthologs creating a set of *G. niphandrodes* genes that are likely oocyst-expressed kinases. The full strategy can be found here: <http://cryptodb.org/cryptodb/im.do?s=88749e65a931cae9>

1. *Use the text search to find genes whose records contain the word kinase in their records.* As in Subheading 2.2.3, **Step 1** above, initiate a text search for the term “kinase” from the dedicated Gene text search (Fig. 4a). The search returns over 3500 genes across all organisms in CryptoDB. Future versions of CryptoDB may return different results as new genome sequences become available and are integrated, and the annotation of existing genome sequences is updated.
2. *Expand the set of kinases by adding a search for genes that are annotated with the GO term GO:0016301: kinase activity.* GO Terms are part of a controlled vocabulary describing the biological function, molecular process, or location of gene products. Terms are assigned to genes during the annotation process. CryptoDB enhances GO annotations by running InterProScan on all genome sequences and integrating the analysis results. A search for genes annotated with the kinase activity GO term may identify genes with kinase activity that do not contain the word kinase in their records. To extend the strategy with this search, click the *Add Step* button (Fig. 16b) and navigate the *Add Step* panel through *Run a new Search for Genes, Function Prediction, Go Term*. When the GO Term search page appears, begin typing the GO term in the *GO Term or GOID* parameter and then choose the correct item (*GO:0016301: kinase activity*) from the list.

Table 2
Methods for combining search results

Operator	Symbol	Combined result will contain
1 intersect 2		IDs in common between 1 and 2
1 union 2		A nonredundant set of IDs from list 1 and list 2
1 minus 2		IDs unique to 1
2 minus 1		IDs unique to 2
1 relative to 2		IDs whose features are new each other (collocated) in the genome

- Since this search is the second step of a strategy, the method of combining the GO Term search results with the Text term search results must be indicated before running the search (Table 2). Choose *1 Union 2* to produce a Step 2 result that contains all genes from both searches. The GO term search returns over 1200 genes annotated with the GO term kinase activity, and the combined result is a set of over 3500 genes with gene products that likely have kinase activity (Fig. 16c).
3. *Reduce the Step 2 result to only those kinases that are highly expressed in oocysts.* To ensure that the kinases in Step 2 are expressed during the oocyst stage, intersect the Step 2 result (kinases) with a search for genes that are highly expressed in the oocyst. Click *Add Step* (Fig. 16c, arrow 7) and navigate the Add Step panel through *Run a new Search for, Genes, Transcriptomics, RNA Sequence* (similar to Fig. 16b). Notice that *G. niphandrodes* is not represented in the *Organism* column, indicating that CryptoDB does not contain RNA sequence data sets associated with *G. niphandrodes*. However, CryptoDB does offer a data set that contains expression values from the *C. parvum* oocyst and intracellular transcriptomes determined by RNA sequence analysis [27]. Searching these data will return genes expressed in oocyst, and the intersection of the oocyst genes with the Step 2 result will find genes in common between the two lists: kinases that are expressed during the oocyst stage.
- Choose the *Percentile* search for the data set “Transcriptome of oocyst and intracellular stages (Widmer et al.)” (Fig. 17a).

Fig. 17 Example 1 strategy, adding Step 3. (a) RNA sequence evidence table for choosing data set to query. (b) Percentile RNA sequence search page specific to the *Transcriptome of oocyst and intracellular stages* data set. The search is configured to search the Oocyst sample (arrow 1) for genes in the 70–100% percentile (arrow 2) of expression in the sample and intersect (arrow 3) the oocyst expressing genes with the Step 2 result. (c) The three-step strategy result

This search returns genes based on the percentile rank of expression values within a sample. CryptoDB analyzed the raw RNA sequence data to produce expression values and determined the percentile rank of each gene within samples. To configure the search to find genes that are highly expressed in oocysts, choose *Oocysts* for the *Samples* parameter (Fig. 17b, arrow 1) and set the *Minimum expression percentile* to 70 (Fig. 17b, arrow 2). Choose to intersect the results of this new search with Step 2 using *2 Intersect 3* (Fig. 17b, arrow 3). The Step 3 result will have genes that are present in the Step 2 result as well as the expression search and, therefore, have both biological properties, kinase activity, and high expression in oocysts. Click *Run Step* to initiate the expression search and combine the results with Step 2. The expression search returns over 1100 genes that are highly expressed in oocysts and over 40 of these are likely to have kinase activity (Fig. 17c).

4. *Use the Transform by Orthology tool to convert the C. parvum oocyst kinases to their G. niphandrodes orthologs.* This step of the strategy takes advantage of expression data available in *C. parvum* to make inferences about possible expression of *G. niphandrodes* genes. Click *Add Step* (Fig. 17c, arrow 4) and choose *Transform by Orthology* in the first column of the Add Step panel (Fig. 18a). Set the transform to *Gregarina niphandrodes* using the *Filter list below* tool. Begin typing the organism name and select the correct organism from the list (Fig. 18a, arrow 1). Click *Run Step* to initiate the transform. The transform returns over 50 *G. niphandrodes* genes. This final strategy result is a set of *G. niphandrodes* genes that likely proteases expressed in oocysts. The full strategy can be found here: <http://cryptodb.org/cryptodb/im.do?s=88749e65a931cae9>
5. *Explore your results.* A critical review of the results is advisable to ensure credibility. For example, review the *Product Description* column of the *Gene Result* to see if they seem like kinase (Fig. 18b, arrow 2). Use the links in the *Gene ID* column (Fig. 18b, arrow 3) to visit a few gene pages which may contain information about domains associated with kinase activity.
6. *Analyze your results.* Determine functional and metabolic pathway enrichment for the *G. niphandrodes* gene set. CryptoDB offers enrichment analyses that can be accessed from the *Analyze Results* tab (Fig. 18c, arrow 4) of any search result. Focus the strategy on the Step 4 result by clicking on the search box in the strategy panel. The active result is highlighted in yellow and the genes from that result populate the Gene Result table. Click the blue *Analyze Results* tab to create a *New Analysis* tab that will contain the analysis results and choose to run a *Gene Ontology Enrichment* (Fig. 19a). Click *Submit* to run the default analysis (Fig. 19b, arrow 1) which will apply a Fischer's exact test to

compare the biological process GO terms assigned to genes in the gene set to a background consisting of all genes in the genome. The analysis returns GO terms that are statistically enriched, appearing more often in the gene result than in all genes in the genome. The default GO enrichment of the *G. niphandrodes* oocyst expressed kinases returns over 120 enriched GO terms with p-values as low as 1×10^{-28} . The *Open in Revigo* button above the enrichment result table (Fig. 19c, arrow 2) opens the Revigo search page and populates the search with the GO enrichment results. Revigo is a clustering and visualization tool for enrichment analyses (Fig. 19c).

2.4.2 Isolate Strategy Example

Find isolates collected from the feces of humans in Europe that were genotyped with GP60. While Strategy Example 1 returned genes with shared biological properties, this strategy retrieves isolate sequences that share characteristics such as geographic location and isolation source. The sample characteristics (metadata) that CryptoDB downloads with the NCBI Popset isolate sequences are used to assemble a set of isolates that share your desired characteristics. This four-step strategy creates a set of isolates that were collected from the feces of humans in Europe and were typed using the gene encoding GP 40/15. The full strategy can be accessed here: <http://cryptodb.org/cryptodb/im.do?s=2418e5fa79e56fde>

1. *Find Popset Isolates Sequence records that were collected in Europe.* Navigate to *Identify Popset Isolate Sequences based on Geographic Location* from the home page *Search for Other Data Types* panel (Fig. 20a). Set the *Geographic Location* parameter to Europe and click *Get Answer*. The search returns over 1700 Popset sequences known to be collected in Europe.
2. *Filter the European isolates to include only those isolated from feces.* Click *Add Step* in the strategy panel (Fig. 20a, arrow 1) and navigate the panel through *Run a new Search for, Popset Isolate Sequences, Isolation Source* (Fig. 20b). From the *Add Step* search form, set the *Isolation Source* to feces (Fig. 20c). Choose to intersect the results of Step 1, the geographic location search, with the isolation source search so that the Popset Isolates in the Step 2 result have both properties, isolated in Europe from feces. Click *Run Step* to initiate the search. The Isolate Source search returns over 3400 isolate sequence records, and the Step 2 result contains over 600 isolates that were isolated in Europe from feces (Fig. 20c).

Fig. 20 Example strategy 2, Steps 1 and 2. (a) Home panel showing the link for the search that returns Popset Isolate records based on *Genomic Location*. Search page shown configured to return isolates collected in Europe. Strategy panel depicting results of the search. (b) *Add Step* panel showing navigation to the *Popset Isolate by Isolation Source* search. (c) *Isolate Source* search form in the *Add Step* panel configured to search for Popset Isolates that were isolated from feces and intersect that result with the Step 1 result

3. *Filter the European isolates from feces to include only those genotyped with sporozoite antigen gp40/15.* Click *Add Step* in the strategy panel (Fig. 20c, arrow 2) and navigate through *Run a new Search for, Popset Isolate Sequences, Locus Sequence Name*. Set the *Locus Sequence Name* parameter to sporozoite antigen GP40/15 (an alias for GP60) (Fig. 21a) and choose *2 Intersect* to combine the Locus Sequence search with the Step 2 result before clicking *Run Step*. The new search returns over 2600 isolates and the intersection with Step 2 returns over 180 isolates that were isolated from feces in Europe and genotyped with sporozoite antigen GP40/15.
4. *Filter the Step 3 result by host to include only sequences isolated from human.* Click *Add Step* after the Step 3 result (Fig. 21a, arrow 1) and navigate the *Add Step* panel through *Run a new Search for, Popset Isolate Sequences, Host Taxon*. To choose only *Homo sapiens* for the *Host* parameter, begin typing *Homo sapiens* in the search box and then click *Select Only These* when the tree is reduced to the correct choice. Choose *3 Intersect* the Host Taxon search with the Step 3 result and click *Run Step* to initiate the search. The search returns over 2200 Popset isolate sequences from human and the final strategy result returns 95 Popset isolate sequences that were genotyped with sporozoite antigen GP40/15 and isolated from the feces of humans in Europe. The full strategy can be accessed here: <http://cryptodb.org/cryptodb/im.do?s=2418e5fa79e56fde>
5. *Explore and analyze your results by running a multiple sequence alignment for the Croatian and Czech Republic isolates.* Click on the *Popset Isolates Geographic Location* tab to see a map of the collection sites for isolates in the result (Fig. 21b, inset). Pin color represents the number of isolates from that location. Return to the *Popset Isolate Sequences* tab and sort the Geographic location column to place the isolates from Croatia and Czech Republic at the top (Fig. 21b, arrow 2). Click the box next to the Popset Sequence IDs for the two Croatian isolates and scroll to the bottom of the list to click *Run ClustalW on Checked Strains*. The alignment appears in a new window (Fig. 21c, left side). The sequences appear to be very similar with few differences noted in pink. Return to the *Popset Isolate Sequences* result and run the multiple sequence alignment

Fig. 21 (continued) strategy result showing the addition of the *Locus Sequence Name* results. **(b)** Final strategy result shown in *My Strategies* including an expansion of the *Popset Isolate Sequences Geographical Location* tab which displays a map of isolates in geographic space (inset). Popset Isolate Sequences result with sortable columns (arrow 2) and a built in sequence alignment tool to easily return isolate ClustalW alignments of checked Popset Sequence IDs. **(c)** Example of alignment results between Croatian isolates (left) and Croatian plus Czech Republic isolates. Pink highlights indicate sequence divergence

on the three Czech Republic isolates in addition to the Croatian isolates. The results reveal (Fig. 21c, right side) differences between the groups of isolates. The Croatian isolates are similar to each other as are the Czech Republic isolates. However, many differences are noted between the Croatian and Czech Republic isolates.

2.5 Data Analysis of Your Sequence Data

Sequence alignment and enrichment analysis tools are an integral part of CryptoDB record pages and strategy results. Comparative genomics often relies on sequence alignments to reveal differences between strains or isolates. For easy access, tools for aligning gene sequence from annotated genome sequences (Fig. 9a) or isolate sequence in the region of the gene (Fig. 10a, b) are integrated into the gene record page. Popset Isolate sequence alignments can be performed from any isolate search or strategy result page (Fig. 21b, c).

While the strategy system facilitates reducing large volumes of data to meaningful gene sets, enrichment analyses support interpretation by identifying over-represented functional annotations such as GO Terms or metabolic pathways in a gene set. The CryptoDB enrichment analyses apply a Fischer's exact test to compare annotations assigned to the genes in a search result to the entire genome. Tools to determine gene ontology term, metabolic pathway, and gene product description term enrichment are readily accessible from a result page under the blue *Analyze Results* tab on any gene search or strategy result page (Fig. 18b, arrow 3 and Fig. 19).

CryptoDB also offers an environment for analyzing raw large-scale data such as for differential expression from RNA sequence or variant calling from whole-genome sequence of isolates. The following section describes how to use the EuPathDB Galaxy data analysis service developed in partnership with Globus Genomics to analyze your own high-throughput next-generation sequence data. Galaxy is an analysis tool which is available on all EuPathDB sites.

2.5.1 CryptoDB/ EuPathDB Galaxy Data Analysis Service

Galaxy is an open-source web-based platform for data intensive research that reduces the barrier of access to bioinformatic analysis since command-line computing is not required. A variety of bioinformatics tools are available and can be used one at a time or dropped into a graphic workspace and chained together to create workflows. Workflows and their results can be shared between users or published to the community. The CryptoDB Galaxy is preloaded with reference genome sequences and offers preassembled workflows for reference-based RNA sequence analysis as well as variant calling (SNP detection). Analysis results are easily ported to CryptoDB for private viewing in the genome browser side-by-side with data already integrated into CryptoDB.

The following example employs a shared EuPathDB workflow to call variants between a diarrheal isolate and the reference genome. A recent study of genetic diversity in Bangladesh performed whole-genome sequencing on 63 *Cryptosporidium hominis* isolates [13]. The published analysis calls single-nucleotide polymorphisms between the *C. hominis* isolates and *C. parvum* IowaII. We will use EuPathDB Galaxy to repeat a portion of this analysis. The data are available in the sequence read archive repositories (Project ID = PRJEB24168 at ENA and SRA) [51, 52] and can easily be imported to a Galaxy account. The workflow performs FastQC [53] to check the isolate's read quality and uses Sickle [54] for adapter trimming, Bowtie2 [55] for aligning the isolate sequences to the reference genome sequence, FreeBayes [56] for calling variants, SNPEff [57] for annotating and predicting the effects of genetic variants, and SNPsift [58] for manipulating the variant call files produced by the workflow.

1. *Visit and explore the CryptoDB Galaxy page by clicking Analyze My Experiment in the header menu bar.* Registration with both Globus Genomics and CryptoDB are required. First-time users will be prompted with a series of screens to create and log into their accounts. Once registered, Galaxy opens to a welcome page in the center panel (Fig. 22a, arrow 1) which contains links to tutorials as well as the EuPathDB preconfigured workflows and is controlled by the center top menu bar (Fig. 22a, red border). The center panel also serves as a workspace for configuring tools, starting and editing workflows, and file management. The tools panel (Fig. 22a, arrow 2) offers bioinformatics tools for a variety of purposes, including importing data into Galaxy, next-generation sequencing applications, data manipulation, and statistical analysis. Choosing a tool from the left panel (Fig. 22a, arrow 3) opens the tool in the center panel (Fig. 22b) where one can change default parameter values (Fig. 22b, left) and initiate the analysis. The History panel presents a running history of files imported or created during the analyses (Fig. 22a, arrow 4 and 22b, right). The Tools and History panels can be collapsed (Fig. 22a, arrow 5) creating a larger workspace in the center.
2. *Import isolate sequence data into Galaxy.* To import the isolate sequence data directly from EBI, open the *Get Data via Globus from the EBI server* tool from the tools panel (Fig. 22a, arrow 3) and enter SAMEA104459070 for the parameter *Enter your ENA Sample ID*. In this case, the sample ID is provided but retrieving Sample IDs is easily accomplished by searching the ENA or SRA repositories with a keyword (such as author name) the project ID (PRJEB24168, which is usually documented in the publication) to retrieve the data set record.

A

B

C

Fig. 22 EuPathDB Galaxy, layout, data upload, and workflow. (a) EuPathDB Galaxy home page accessed through the *Analyze My Experiment* link in the gray menu bar. The center workspace (arrow 1) is flanked by the *Tools* (arrow 2) and *History* (arrow 4) panels. Shown are the *Get Data via Globus from the EBI server*

Enter parameter values for the *Data type to be transferred* as fastq and the *Single or Paired-Ended* as paired (Fig. 22b, left). Click execute to initiate the upload. Tiles representing files appear in the history panel and will turn from gray to yellow when the job starts and then to green when the job is finished. The files are named ERR2240057_1 and ERR2240057_2 which represent forward and reverse RNA sequence runs associated with the sample ID.

3. *Run the preconfigured variant calling workflow shared by EuPathDB.* Return to the welcome page by clicking *Analyze Data* in the center panel (Fig. 22a, red border) and choose the *EuPathDB Workflow for Variant Calling, paired-end sequencing* (Fig. 22a, arrow 6). Indicate the *Input data* sets (Fig. 22c ERR2240057_1 for forward reads; ERR2240057_2 for reverse reads). Three tools in the workflow (Bowtie2, FreeBayes and SNPEff) require the reference genome as input. Since all EuPathDB genomes are preloaded into EuPathDB Galaxy, defining the reference genome is as easy as choosing it from a pick list. Scroll down carefully to change the reference genome to *CryptoDB-37_CparvumIowaII_Genome* in three places (Fig. 22c, arrow 7). Click *Run workflow* (Fig. 22c, arrow 8) to start the workflow. History files appear and will indicate successful completion of a step when green or failure when red.
4. *Explore the results.* The history panel contains a tile for each output file created by the workflow (Fig. 23a). Some files are hidden from the history panel but can be shown by clicking the *hidden* link (Fig. 23a, arrow 1). Clicking the name of the file reveals file attributes within the history tile, whereas clicking the eye icon populates the center panel with file contents. Files can also be downloaded for use in local programs such as excel (Fig. 23a, red border).
5. *View variants in CryptoDB GBrowse.* The FreeBayes output vcf file (variant call file) can easily be converted to a graphic file (bigwig) that can be exported to CryptoDB for viewing. Click the edit icon next and open the Convert Format tab that appears in the center panel (Fig. 23b). Choose *Convert BED, GFF, or VCF to BigWig* and click *Convert*. A new history tile representing the new file will appear. Once the file conversion is complete, use the *EuPathDB Export Tool, Bigwig Files to EuPathDB*, to transfer the file to the *My Data Sets* section of CryptoDB (Fig. 23c). This tool requires an entry for all five

←
Fig. 22 (continued) tool (arrow 3) for uploading data from EBI to your EuPathDB Galaxy account and the link for the preconfigured workflow for variant calling (arrow 6). **(b)** *Get Data via Globus from the EBI server* tool (left) and a completed upload in the history panel (right). **(c)** Variant calling workflow showing parameters for choosing input data sets (arrow 7) and the dropdown menu for selecting a reference genome. The *Run Workflow* button (arrow 8) for initiating the workflow

A

```
##FastQC      0.11.3
>>Basic Statistics      pass
#Measure      Value
Filename      ERR2240057_1.fastq.gz
File type      Conventional base calls
Encoding      Sanger / Illumina 1.9
Total Sequences 764856
Sequences flagged as poor quality      0
Sequence length 150
%GC      30
>>END_MODULE
>>Per base sequence quality      pass
#Base      Mean      Median      Lower Quartile      Upper Quartile      10th
Percentile      90th Percentile
1      33.55794685535578      34.0      33.0      34.0      32.0
35.0
2      33.725667315154745      34.0      33.0      34.0      32.0
35.0
3      33.795523601828314      34.0      33.0      35.0      33.0
35.0
4      33.822008325750204      34.0      33.0      35.0      33.0
35.0
5      33.760668674887825      34.0      33.0      34.0      33.0
35.0
6      36.43706789251833      37.0      37.0      37.0      34.0
37.0
7      36.43471058604496      37.0      37.0      37.0      34.0
```

History

MiMB Crypto2
11 shown, 7 hidden
1.31 GB

ata 2 and data 1

4: FastQC on data 1: RawData
1,444 lines
format: txt, database: ?

##FastQC 0.11.3
>>Basic Statistics pass
#Measure Value
Filename ERR2240057_1.fastq.gz
File type Conventional base call
Encoding Sanger / Illumina 1.9

3: FastQC on data 1: Web page

B

Attributes Convert Format Datatype Permissions

Convert to new format

Convert BED, GFF, or VCF to BigWig ▾

This will create a new dataset with the contents of this dataset converted to a new format.

Convert

History

13: FreeBayes on data 12 (variants)
~200,000 lines
format: vcf, database: CryptoDB-37_Cparvumlowall_Genome

display with IGV local

1.Chrom 2.Pos

##fileformat=VCFv4.1

C

Tools

search tools

Get Data

EUPATHDB APPLICATIONS

EuPathDB Export Tools

Bigwig Files to EuPathDB Export one or more bigwig files to EuPathDB where they can be viewed as tracks in the Genome Browser.

Fig. 23 EuPathDB Galaxy, results and file conversion. (a) History panel actions for retrieving file information (click the Filename) or for displaying the file contents in the center panel (click the eye icon) and for file download (red border). (b) Tool for converting text files with genomic location to graphic bigwig files which can be displayed in the genome browser. Access the conversion tool through the Edit icon (pencil) in the history panel. (c) EuPathDB Export Tools section showing the *Bigwig Files to EuPathDB* tool

Bigwig Files to EuPathDB Export one or more bigwig files to EuPathDB where they can be viewed as tracks in the Genome Browser. (Galaxy Tool Version 1.0.0) Options

My Data Set name:

 The export will create a new My Data Set in a EuPathDB site. Specify the name for the new My Data Set.

Bigwig files:

20: Convert BED_GFF_or_VCF _to_BigWig_on_data_13
 18: SnpSift Filter on data 16
 16: SnpEff on data 15
 14: BAM to BigWig on data 12
 12: Sort on data 11: sorted BAM

 Select the files to include in the new EuPathDB My Data Set. The bigwig files you select here must be mapped to the reference genome that you select below.

Reference genome:

 The bigwig files you selected above must be mapped to the reference genome that you select here.

My Data Set summary:

My Data Set description:

Fig. 24 Bigwig files to EuPathDB tool. Tool is shown configured to create a data set called *C hominis* isolate which will contain one Bigwig file that is associated with the *C. parvum* reference genome. All five parameters need entries, or the tool will return a request for parameter values

parameters before executing. Open the export tool and make an entry for all parameters, choose the file to transfer, and make an entry for all parameters, *My Data Set name*, *Bigwig files*, *Reference genome*, *My Data Set summary*, and *My Data Set description* (Fig. 24). A history tile representing the export will appear. Once the export is finished and your history tile is green, visit CryptoDB *My Data Sets* (Fig. 25a) page to install the file into GBrowse. The *My Data Sets* page tabulates all private data sets that are shared with or uploaded by you. Open the record for the new data set by clicking the data set name in the *My Data Sets* table (Fig. 25a, arrow 1). The *GBrowse Tracks* table contains a *Send to GBrowse* button that installs the file in GBrowse (Fig. 24b) and, upon file installation, becomes a link (*View in GBrowse*) to directly view the file in GBrowse (Fig. 24c). GBrowse is itself a complex analysis tool. Consulting

Fig. 25 My data sets section. (a) *My Data Sets* page, accessed through the gray menu bar, tabulates your personal data sets. (b) GBrowse Tracks table from the *My Data Sets* page with options to send the file to GBrowse for viewing and to *View in GBrowse* after the file has been sent. (c) Variant file (vcf converted to bigwig) displayed in GBrowse

a GBrowse user guide [59] is recommended for assistance with configuring visualizations, especially navigation and adjusting scales when viewing quantitative results.

2.6 Summary

This exciting time for *Cryptosporidium* omics-based research supports biological discovery and a changing view of *Cryptosporidium* as a pathogen. CryptoDB and its associated analysis tools reduce

the threshold of access to valuable omics data and provide a powerful data mining and analysis platform for *Cryptosporidium* researchers and related research communities. CryptoDB is updated roughly every 2 months with new data and new features. Changes made to the database will affect data mining outcomes; thus, strategies should be saved for easy execution when new data are available. As laboratory breakthroughs render *Cryptosporidium* more amenable to advanced molecular and phenotypic analyses such as genetic crosses, high-throughput phenotypic screens, and CRISPR-Cas, additional tools developed for other organisms represented in the EuPathDB family of databases will be provisioned for CryptoDB.

References

1. NIAID-Funded Bioinformatics Resource Centers (2018). <https://www.niaid.nih.gov/research/bioinformatics-resource-centers>
2. Eukaryotic Pathogen Database Resources (EuPathDB) (2018). <https://eupathdb.org/eupathdb/>
3. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Gruning BA, Guerler A, Hillman-Jackson J, Hiltmann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 46(W1):W537–W544. <https://doi.org/10.1093/nar/gky379>
4. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 15(10):1451–1455. <https://doi.org/10.1101/gr.4086505>
5. Popset (2018). <https://www.ncbi.nlm.nih.gov/popset>
6. Gregarina niphandroides genome sequencing project (2018). <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA59799/>
7. Templeton TJ, Pain A (2016) Diversity of extracellular proteins during the transition from the ‘proto-apicomplexan’ alveolates to the apicomplexan obligate parasites. *Parasitology* 143(1):1–17. <https://doi.org/10.1017/S0031182015001213>
8. Reese ML, Zeiner GM, Sacij JP, Boothroyd JC, Boyle JP (2011) Polymorphic family of injected pseudokinases is paramount in *Toxoplasma* virulence. *Proc Natl Acad Sci U S A* 108(23):9625–9630. <https://doi.org/10.1073/pnas.1015980108>
9. GenBank (2018). <https://www.ncbi.nlm.nih.gov/genbank/>
10. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. *Nucleic Acids Res* 41(Database issue):D36–D42. <https://doi.org/10.1093/nar/gks1195>
11. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13(9):2178–2189
12. Feng Y, Li N, Roellig DM, Kelley A, Liu G, Amer S, Tang K, Zhang L, Xiao L (2017) Comparative genomic analysis of the IId subtype family of *Cryptosporidium parvum*. *Int J Parasitol* 47(5):281–290. <https://doi.org/10.1016/j.ijpara.2016.12.002>
13. Gilchrist CA, Cotton JA, Burkey C, Arju T, Gilmartin A, Lin Y, Ahmed E, Steiner K, Alam M, Ahmed S, Robinson G, Zaman SU, Kabir M, Sanders M, Chalmers RM, Ahmed T, Ma JZ, Haque R, Faruque ASG, Berriman M, Petri WA (2018) Genetic diversity of *Cryptosporidium hominis* in a Bangladeshi community as revealed by whole-genome sequencing. *J Infect Dis* 218(2):259–264. <https://doi.org/10.1093/infdis/jiy121>
14. Hadfield SJ, Pachebat JA, Swain MT, Robinson G, Cameron SJ, Alexander J, Hegarty MJ, Elwin K, Chalmers RM (2015) Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. *BMC Genomics* 16:650. <https://doi.org/10.1186/s12864-015-1805-9>
15. Widmer G, Lee Y, Hunt P, Martinelli A, Tolckoff M, Bodi K (2012) Comparative

- genome analysis of two *Cryptosporidium parvum* isolates with different host range. *Infect Genet Evol* 12(6):1213–1221. <https://doi.org/10.1016/j.meegid.2012.03.027>
16. Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Res* 28(1):304–305
 17. UniProt C (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 40(Database issue):D71–D75. <https://doi.org/10.1093/nar/gkr981>
 18. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>
 19. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. *Nucleic Acids Res* 32:D138–D141
 20. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SC, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res* 45(D1):D190–D199. <https://doi.org/10.1093/nar/gkw1107>
 21. Letunic I, Bork P (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* 46(D1):D493–D496. <https://doi.org/10.1093/nar/gkx922>
 22. Marzano V, Mancinelli L, Bracaglia G, Del Chierico F, Vernocchi P, Di Girolamo F, Garrone S, Tchidjou Kuekou H, D'Argenio P, Dallapiccola B, Urbani A, Putignani L (2017) “Omic” investigations of protozoa and worms for a deeper understanding of the human gut “parasitome”. *PLoS Negl Trop Dis* 11(11):e0005916. <https://doi.org/10.1371/journal.pntd.0005916>
 23. Kissinger JC, Hermetz KE, Woods KM, Upton SJ (2018) Enrichment of *Cryptosporidium parvum* from in vitro culture as measured by total RNA and subsequent sequence analysis. *Mol Biochem Parasitol* 220:5–9. <https://doi.org/10.1016/j.molbiopara.2017.12.004>
 24. dbEST (2018). <https://www.ncbi.nlm.nih.gov/dbEST/>
 25. Sequence Read Archive (2018). <https://www.ncbi.nlm.nih.gov/sra/>
 26. Isaza JP, Galvan AL, Polanco V, Huang B, Matveyev AV, Serrano MG, Manque P, Buck GA, Alzate JF (2015) Revisiting the reference genomes of human pathogenic *Cryptosporidium* species: reannotation of *C. parvum* Iowa and a new *C. hominis* reference. *Sci Rep* 5:16324. <https://doi.org/10.1038/srep16324>
 27. Mirhashemi ME, Noubary F, Chapman-Bonofiglio S, Tzipori S, Huggins GS, Widmer G (2018) Transcriptome analysis of pig intestinal cell monolayers infected with *Cryptosporidium parvum* asexual stages. *Parasit Vectors* 11(1):176. <https://doi.org/10.1186/s13071-018-2754-3>
 28. Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michalek J, Saxena A, Shanmugam D, Tayyrov A, Veluchamy A, Ali S, Bernal A, del Campo J, Cihlar J, Flegontov P, Gornik SG, Hajduskova E, Horak A, Janouskovec J, Katris NJ, Mast FD, Miranda-Saavedra D, Mourier T, Naeem R, Nair M, Panigrahi AK, Rawlings ND, Padron-Regalado E, Ramaprasad A, Samad N, Tomcala A, Wilkes J, Neafsey DE, Doerig C, Bowler C, Keeling PJ, Roos DS, Dacks JB, Templeton TJ, Waller RF, Lukes J, Obornik M, Pain A (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife* 4:e06974. <https://doi.org/10.7554/eLife.06974>
 29. Zhang H, Guo F, Zhou H, Zhu G (2012) Transcriptome analysis reveals unique metabolic features in the *Cryptosporidium parvum* oocysts associated with environmental survival and stresses. *BMC Genomics* 13:647. <https://doi.org/10.1186/1471-2164-13-647>
 30. Mauzy MJ, Enomoto S, Lancto CA, Abrahamsen MS, Rutherford MS (2012) The *Cryptosporidium parvum* transcriptome during in vitro development. *PLoS One* 7(3):e31715. <https://doi.org/10.1371/journal.pone.0031715>
 31. Rowe JS, Shah SS, Motlhagodi S, Bafana M, Tawanana E, Truong HT, Wood SM, Zetola NM, Steenhoff AP (2010) An epidemiologic review of enteropathogens in Gaborone, Botswana: shifting patterns of resistance in an HIV endemic region. *PLoS One* 5(6):e10924
 32. Madrid-Aliste CJ, Dybas JM, Hogue Angeletti R, Weiss LM, Kim K, Simon I, Fiser A (2009)

- EPIC-DB: a proteomics database for studying Apicomplexan organisms. *BMC Genomics* 10(1):38
33. Snelling WJ, Lin Q, Moore JE, Millar BC, Tosini F, Pozio E, Dooley JSG, Lowery CJ (2007) Proteomics analysis and protein expression during sporozoite excystation of *Cryptosporidium parvum* (Coccidia, Apicomplexa). *Mol Cell Proteomics* 6(2):346–355. <https://doi.org/10.1074/mcp.M600372-MCP200>
 34. Sanderson SJ, Xia D, Prieto H, Yates J, Heiges M, Kissinger JC, Bromley E, Lal K, Sinden RE, Tomley F, Wastling JM (2008) Determining the protein repertoire of *Cryptosporidium parvum* sporozoites. *Proteomics* 8(7):1398–1414. <https://doi.org/10.1002/pmic.200700804>
 35. Kanehisa M (2002) The KEGG database. *Novartis Found Symp* 247:91–101. discussion 101–103, 119–128, 244–152
 36. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34(Database issue):D354–D357
 37. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44(D1):D457–D462. <https://doi.org/10.1093/nar/gkv1070>
 38. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Karp PD (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 44(D1):D471–D480. <https://doi.org/10.1093/nar/gkv1164>
 39. Karp PD, Riley M, Paley SM, Pellegrini-Toole A (2002) The MetaCyc Database. *Nucleic Acids Res* 30(1):59–61
 40. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 32(Database issue):D438–D442
 41. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han LY, He JE, He SQ, Shoemaker BA, Wang JY, Yu B, Zhang J, Bryant SH (2016) PubChem substance and compound databases. *Nucleic Acids Res* 44(D1):D1202–D1213. <https://doi.org/10.1093/nar/gkv951>
 42. Welcome to the EuPathDB Galaxy Site (2018). <https://eupathdb.globusgenomics.org/>
 43. Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, Otto TD (2016) Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res* 44(W1):W29–W34. <https://doi.org/10.1093/nar/gkw292>
 44. UniProt C (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43(Database issue):D204–D212. <https://doi.org/10.1093/nar/gku989>
 45. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* 304(5669):441–445. <https://doi.org/10.1126/science.1094786>
 46. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA (2004) The genome of *Cryptosporidium hominis*. *Nature* 431(7012):1107–1112. <https://doi.org/10.1038/nature02977>
 47. Striepen B, Pruijssers AJP, Huang JL, Li C, Gubbels MJ, Umejiego NN, Hedstrom L, Kissinger JC (2004) Gene transfer in the evolution of parasite nucleotide biosynthesis. *Proc Natl Acad Sci U S A* 101(9):3154–3159
 48. Shanmugasundram A, Gonzalez-Galarza FF, Wastling JM, Vasieva O, Jones AR (2013) Library of apicomplexan metabolic pathways: a manually curated database for metabolic pathways of apicomplexan parasites. *Nucleic Acids Res* 41(Database issue):D706–D713. <https://doi.org/10.1093/nar/gks1139>
 49. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD (2016) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32(2):309–311. <https://doi.org/10.1093/bioinformatics/btv557>
 50. Mogi T, Kita K (2010) Diversity in mitochondrial metabolic pathways in parasitic protists *Plasmodium* and *Cryptosporidium*. *Parasitol Int* 59(3):305–312
 51. Takashima Y, Xuan X, Kimata I, Iseki M, Kodama Y, Nagane N, Nagasawa H, Matsumoto Y, Mikami T, Otsuka H (2003) Recombinant bovine Herpesvirus-1 expressing p23 protein of *Cryptosporidium parvum* induces neutralizing antibodies in rabbits. *J Parasitol* 89(2):276–282
 52. European Molecular Biology Laboratory (2018) European Nucleotide Archive. <https://www.ebi.ac.uk/ena>

53. Bioinformatics B (2018) FastQC. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
54. Core UDB (2018) Software: Sickie. <http://bioinformatics.ucdavis.edu/research-computing/software/>
55. University JH (2018) Bowtie 2: Fast and sensitive read alignment. <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>
56. Erik Garrison GM (2012) Haplotype-based variant detection from short-read sequencing. <https://arxiv.org/abs/1207.3907>
57. Cingolani P (2018) SnpEff Documentation. http://snpeff.sourceforge.net/SnpEff_manual.html
58. Cingolani P (2018) SnpSift: filter and manipulate annotated files. <http://snpeff.sourceforge.net/SnpSift.html#filter>
59. Stein LD (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief Bioinform* 14(2):162–171. <https://doi.org/10.1093/bib/bbt001>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.