

Exploring the Gene Page

Note: this exercise uses **ToxoDB** (<https://ToxoDB.org>) as an example database, but the same functionality is available on all VEuPathDB resources.

What is a Gene Page on VEuPathDB?

- The gene page conveniently consolidates links to all the data available on VEuPathDB for a particular gene into a single page

Learning objectives

- Become familiar with the information in gene pages
- Navigate to and from the gene pages
- Use the contents section of the gene page
- Interact with gene page subsections

1. **Navigate to a Gene page:** For this exercise visit the gene page for TGME49_222020 (phosphoglycerate kinase PGKII). How did you get to this gene page? (hint: use the site search to retrieve the gene page link, then click on the gene ID in the results.)

The screenshot shows the ToxoDB website interface. At the top, there is a header with the ToxoDB logo, release information (Release 68, 7 May 2024), and a search bar containing 'TGME49_222020'. Below the header is a navigation menu with links: My Strategies, Searches, Tools, My Workspace, Data, About, Help, and Contact Us. The main content area is titled 'Genes matching TGME49_222020' and shows '1 - 1 of 1' results. On the left, there is a 'Filter results' sidebar with sections for 'Filter Gene fields' and 'Filter organisms'. The 'Filter Gene fields' section has checkboxes for 'Gene ID' and 'Names, IDs, and aliases', both of which are selected. The 'Filter organisms' section has a search bar and checkboxes for 'Sarcocystidae' and 'Toxoplasma', both of which are selected. The main results area shows a single gene entry: 'Gene - TGME49_222020 phosphoglycerate kinase PGKII'. The entry includes the following information: 'Gene name or symbol: PGKII', 'Gene type: protein coding gene', and 'Organism: Toxoplasma gondii ME49'. Below this information, there is a link to 'Fields matched: Gene ID; Names, IDs, and aliases'.

2. Explore the layout of the page

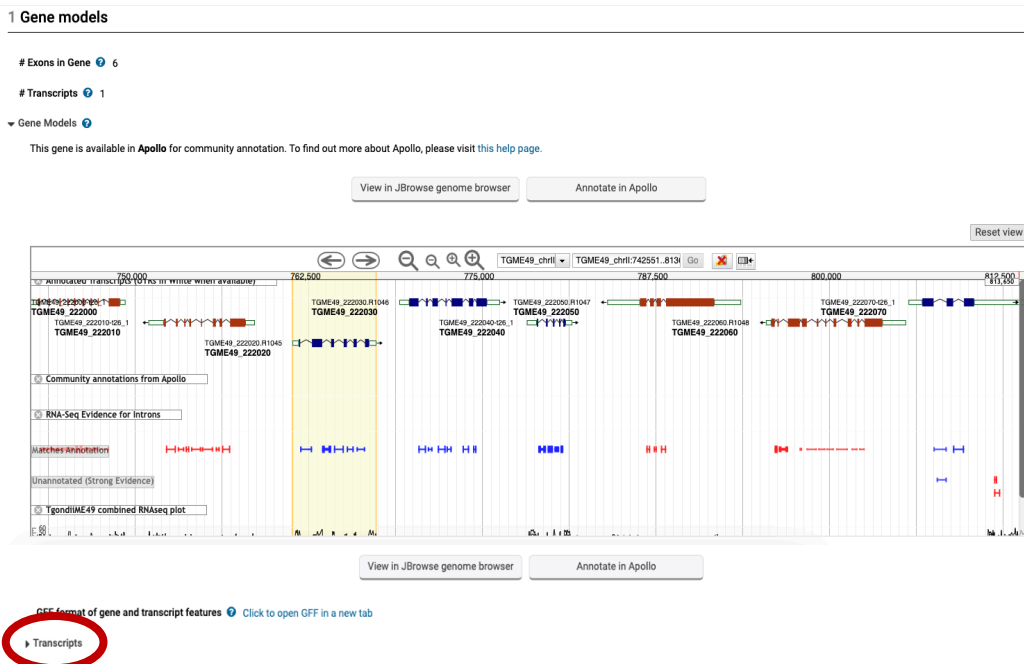
- What information is in the top left (section A in image below)?
 - Can you easily find which chromosome this gene is located on?
 - Is this gene protein coding?
- What do the shortcuts do? (section B)
- Where can you find the contents of this page? (Section C)
- Scroll down and examine what section D contains.

The screenshot shows the Ensembl genome browser interface for the gene **TGME49_222020**, which is a phosphoglycerate kinase PGKII. The page is divided into several sections:

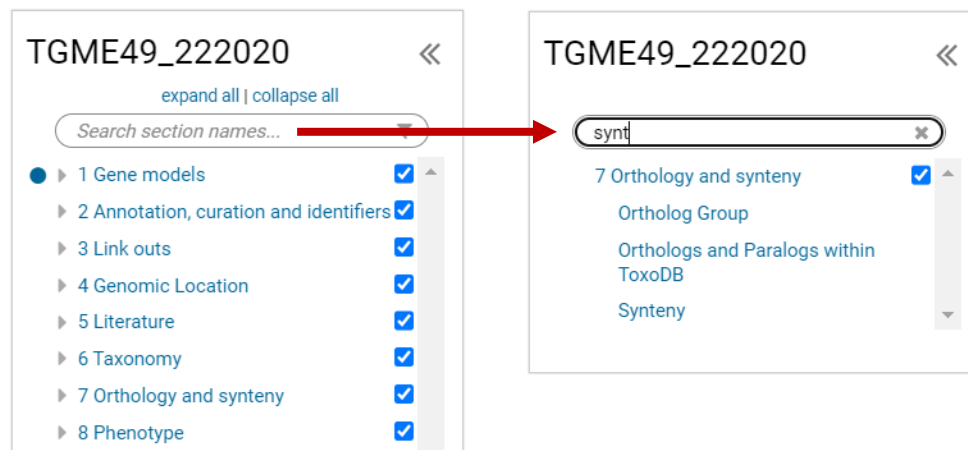
- Section A (Top Left):** Contains gene details such as Name (PGKII), Gene Type (protein coding gene), Biotype Classification (protein_coding), Chromosome (II), Location (TGME49_chrl:761,352..767,399(+)), Species (*Toxoplasma gondii*), Strain (ME49), and Status (Reference Strain).
- Section B (Shortcuts):** A row of icons for Synteny, Alignments, Phenotype, SNPs, Transcriptomics, Protein Features, and Proteomics. Below these icons is a link: "Also see TGME49_222020 in the Genome Browser or Protein Browser".
- Section C (Table of contents):** A sidebar on the left with a search bar and a list of sections: 1 Gene models, 2 Annotation, curation and identifiers, 3 Link outs, 4 Genomic Location, 5 Literature, 6 Taxonomy, 7 Orthology and synteny, 8 Phenotype, 9 Genetic variation, 10 Transcriptomics, and 11 Sequences. Each item has a checkbox to its right.
- Section D (Gene models):** The main content area showing the gene model. It displays "# Exons in Gene" as 6 and "# Transcripts" as 1. Below this, it says "Gene Models" and provides a link to "this help page". At the bottom, there are two buttons: "View in JBrowse genome browser" and "Annotate in Apollo".

3. Explore the gene model section. Scroll down to the gene model section of the gene page.

- What direction is the transcript relative to the chromosome?
- Does the gene have UTRs?
- How many exons does the gene have?
- Does this gene have any available community annotation?
- How long is the transcript? You can determine transcript length by expanding the Transcripts table.



4. **Content navigation.** How do you navigate to the different sections of the page? Use the “Contents” navigation pane on the left side of the page. Gene page content is organized by data type and the section titles serve as links to data within the page. When expanded, each section reveals more navigation links. The content menu can also be filtered using the search function as shown below. Begin typing the ‘synteny’ in the filter to collapse the content menu.



- Click to navigate to the synteny section. Does *Cyclospora cayetanensis* isolate NF1_C8 share synteny in this region?
- Navigate to the Transcript Expression table and open the row for the experiment called “mRNA Illumina sequences from 4 different mouse cell types infected with *T. gondii* and aligned to the ME49 Genome.” (Notice the filter box at the top of the table)
 - What data type was used to produce this data?

- ii. In what cell type does this gene have the highest expression level?
- iii. Open the Data table for this experiment. What is the TPM expression value for unique reads mapped to the Neurons?

- 5. Run an alignment of selected protein sequences.** The Orthologs and Paralogs in ToxoDB table shows all orthologs for TGME49_222020. This table is also a tool for running Clustal Omega on protein, CDS and genomic sequence for this gene and the orthologs you choose from the table. Within all VEuPathDB sites, OrthoMCL is used to determine orthology.
- a. Expand the “Orthologs and Paralogs in ToxoDB” section.
 - b. Select a few genes from the table using the checkbox.
 - c. Scroll to the bottom of the table and click on the Run Clustal Omega button.

<input type="checkbox"/>	TGVEG_222020	phosphoglycerate kinase PGKII	Toxoplasma gondii VEG	no	yes	no
<input checked="" type="checkbox"/>	TGVEG_318230	phosphoglycerate kinase PGKI	Toxoplasma gondii VEG	no	no	no
<input checked="" type="checkbox"/>	TGP89_222020	phosphoglycerate kinase PGKII	Toxoplasma gondii p89	no	yes	no
<input checked="" type="checkbox"/>	TGP89_318230	phosphoglycerate kinase PGKI	Toxoplasma gondii p89	no	no	no

Select sequence type for Clustal Omega multiple sequence alignment:

Please note: selecting a large flanking region or a large number of sequences will take several minutes to align.

☒ Protein
 ☐ CDS (spliced)
 ☐ Genomic

Output format: Mismatches highlighted ▾

- 6. Explore the genetic variation section:** The genetic variation section contains a graphic and a tool for aligning isolate sequences within the region of the gene. Go to the Genetic variation section of the gene page and expand the SNP section. Notice that by default you cannot scroll within the embedded browser window. Scrolling is enabled by default. To reset the image to the default position, choose Reset to Default. To scroll down within the browser window, click and drag or use two-finger scrolling. You can also double click in an area to zoom in.



SNP color code: Dark blue (non-synonymous), light blue (synonymous), Yellow (non-coding), Red (nonsense).

- What kind of SNPs are in this gene?
- Can you see any non-synonymous SNPs?
- How does this compare to the neighboring genes?

7. Explore other sections of the gene page.

Feel free to scroll around the gene page and ask questions for clarification. Here are some questions you may want to ask about this gene:

- Is there evidence that this protein is phosphorylated? (hint: go to the **Proteomics** section and expand the Post Translational Modification section).
- Where is the protein localized? (hint: go to the **Protein Targeting and Localization** section and expand the cellular localization section).
- Is there any phenotypic data available for this gene? What method was used to obtain phenotypic scores? (hint: go to the **Phenotype** section and expand its subsections).
- What other RNA-Seq data available for this gene? (hint: go to the **Transcriptomics** section and expand the subsections called RNA-Seq transcription summary and Transcript Expression).
- Is there transcriptomics data that would provide a high resolution view of cell-to-cell variation? (hint: go to the **Transcriptomics** section and expand the subsections called Single Cell RNA-Seq (scRNA-Seq)).
- Search for “alphaFold” within the section names navigation pane. What data can you obtain in this section? (hint: it will take you to the **Structure Analysis** section)