The VEuPathDB Study Explorer

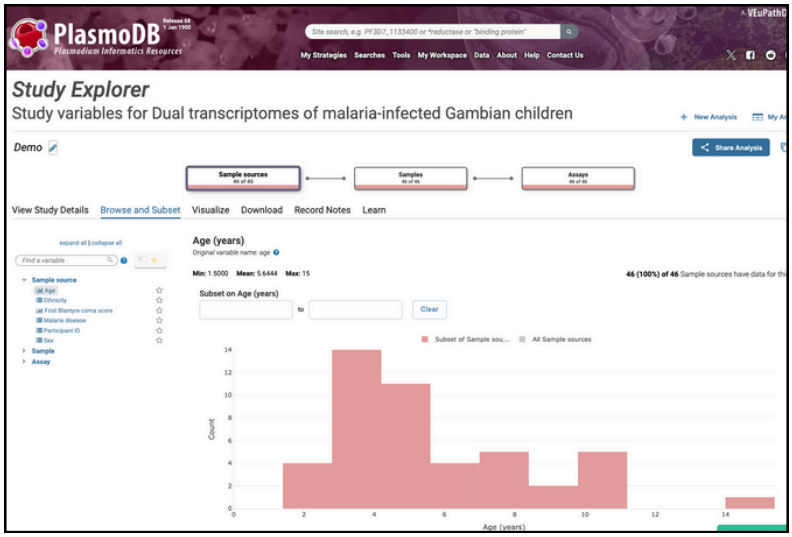
Exploring clinical and WGCNA variables for dual transcriptomes of malaria-infected Gambian children

The **VEuPathDB Study Explorer** is an interactive feature that allows you to

- Learn more about a dataset
- Explore all the variables in a dataset
- Perform exploratory data analysis to visualize associations between two or more variables
- Download the data and work with it on your own

This tutorial describes the features of the study explorer for the WGCNA dataset of 46 malaria-infected Gambian children. This **Study Explorer** offers metadata filters and visualization tools to explore metadata from dual (host and parasite) transcriptomic analysis of Gambian children infected with either severe malaria or uncomplicated malaria.

The Study Explorer can be used, for instance, to help conceptualize co-expression networks and choose a host or parasite module of interest.

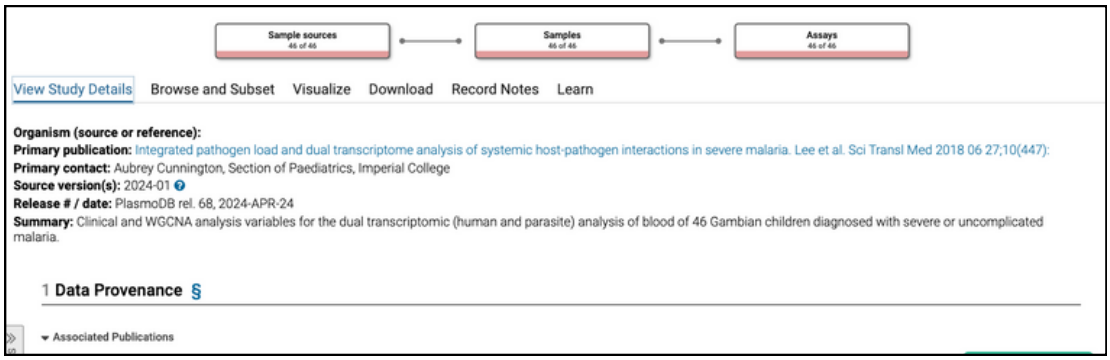


Essential terminology for understanding this dataset

- **WGCNA module**: Whole genome coexpression network analysis (WGCNA) identifies groups of genes that have similar expression across samples. The analysis considers the whole transcriptome data set and clusters coexpressed genes into groups called modules. The modules in VEuPathDB were found using an iterative Weighted Gene Correlation Network Analysis.
- **Eigengene**: The eigengene is created in the WGCNA analysis and is an imaginary gene whose expression profile represents an average gene within the module. Each module is represented by an eigengene. It is very likely that no gene within the module has the same expression profile as the eigengene.

Dataset Diagram

Across the top of the page is a diagram that summarizes the structure of the dataset and the sample size. This dataset contains 46 sample sources, i.e., study participants, and 46 samples, one from each study participant.



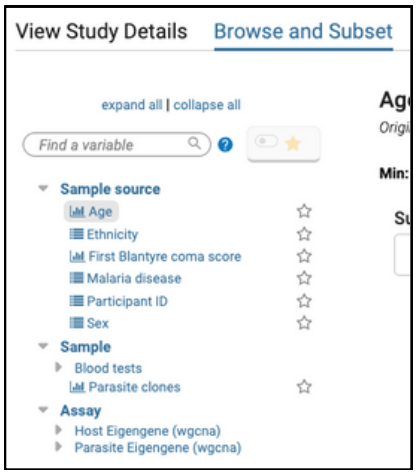
Below the dataset diagram are several tabs.

"View Study Details" tab

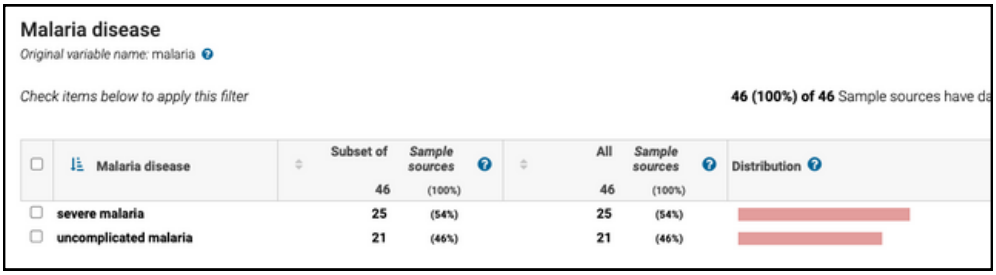
Provides a summary of the dataset, links to associated publications, and a list of study investigators

"Browse and Subset" tab

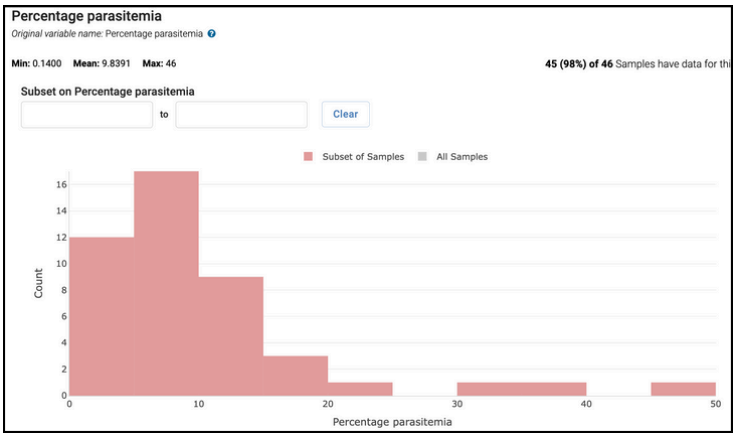
1. Browse through a **hierarchical variable tree**, a list of all the variables in the dataset that is displayed in on the left of the page. In this dataset, you can see variables associated with
   - Sample Source (study participants), such as Age, Sex, and Malaria disease
   - Sample (blood sample) such as Percentage parasitemia
   - Assay (WGCNA) such as host and parasite eigengenes, representing co-expression networks/clusters/modules



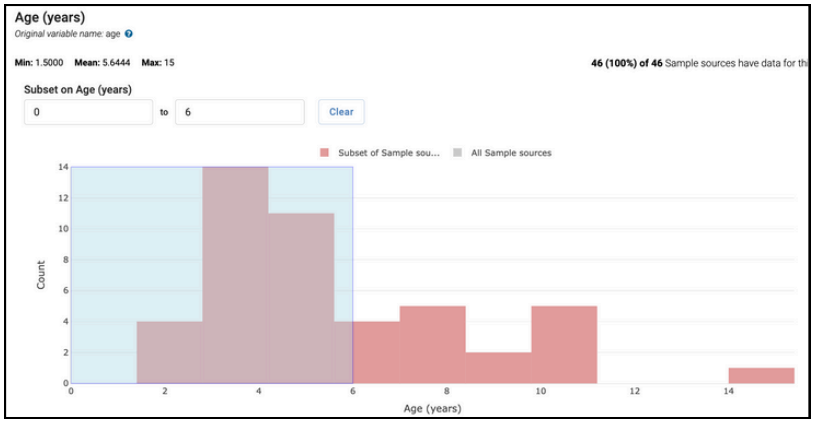
2. View the **univariate distributions** of each of the variables by clicking on the variable label. For example, clicking on Sample Source > *Malaria disease* (a **categorical variable**) displays a frequency table indicating that 25 (54%) of the study participants had severe malaria



Clicking on Sample > *Percentage parasitemia* (a **continuous variable**) displays a histogram, indicating, for instance, that 11 participants had 0-5% parasitemia



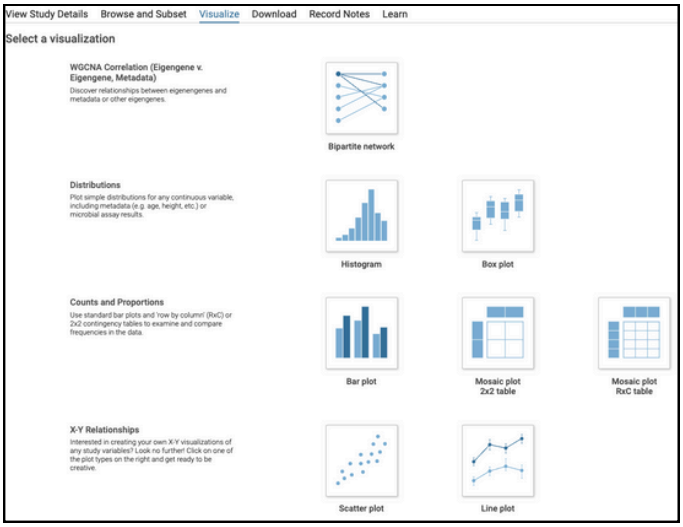
3. **Subset the data** to select observations of interest
For example, to restrict the analysis to participants 6 years old or younger, open the distribution for *Age* and use the 'Subset on Age (years)' section to define the age range: 0 to 6



"**Visualize**" tab

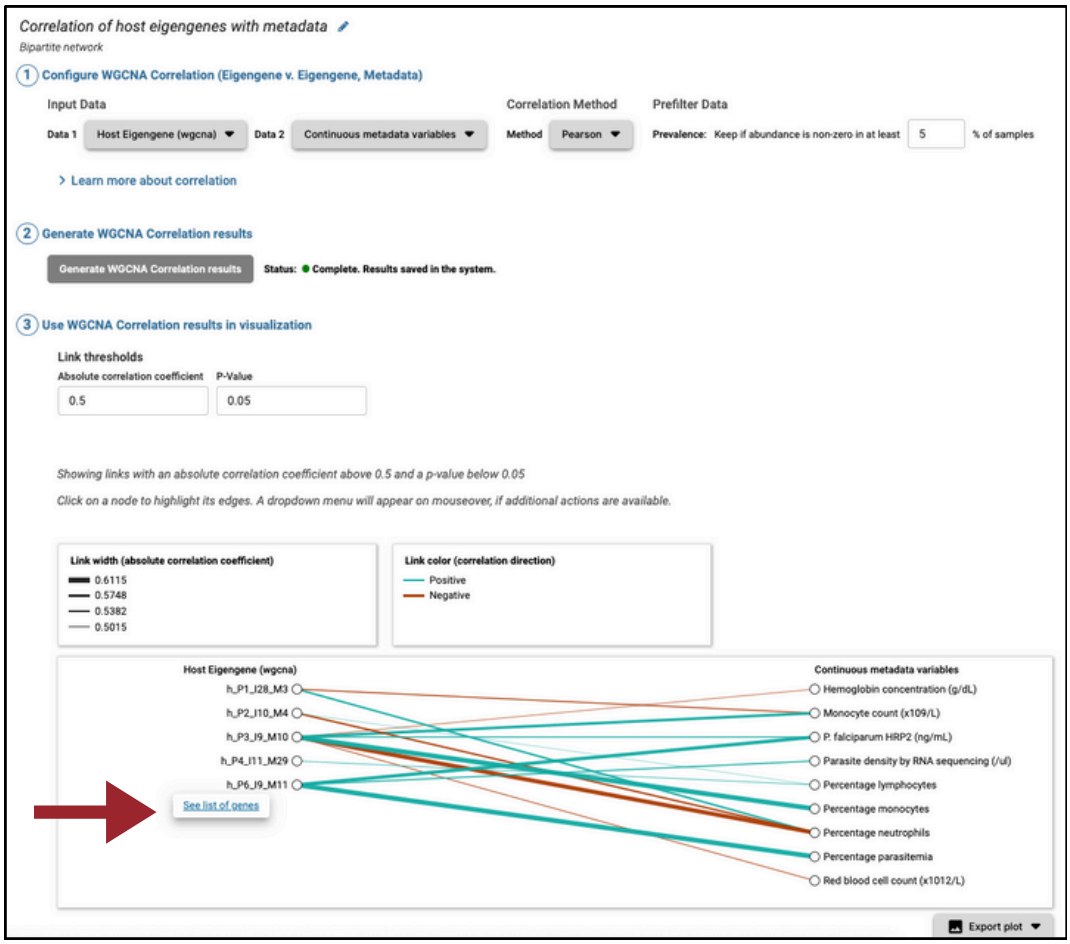
The "Visualize" tab enables you to create graphs and plots to explore associations between two or more variables.

Clicking "New visualization" opens a menu of visualization apps; click on any icon to open the app and configure it.

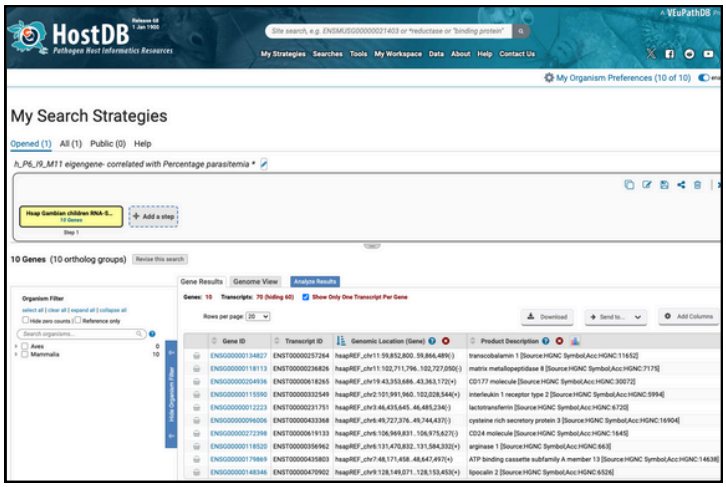


Three Examples of Visualizations

1. Correlation. For example, in this dataset you may want to visualize a bipartite network of host eigengenes and percentage parasitemia, which is one of the continuous metadata variables. To do this:
    - Click on New visualization > WGCNA Correlation
    - Input the Data 1, choosing Host eigengene (wgcna) from the drop-down menu. For Data 2, select continuous metadata variables
    - Choose a correlation method, e.g., Pearson correlation to identify linear trends
    - Click on “Generate correlation results”
    - Visualize the correlation results. You may observe, for instance, that the host eigengene h_P6_I9_M11 is positively correlated with Percentage parasitemia.

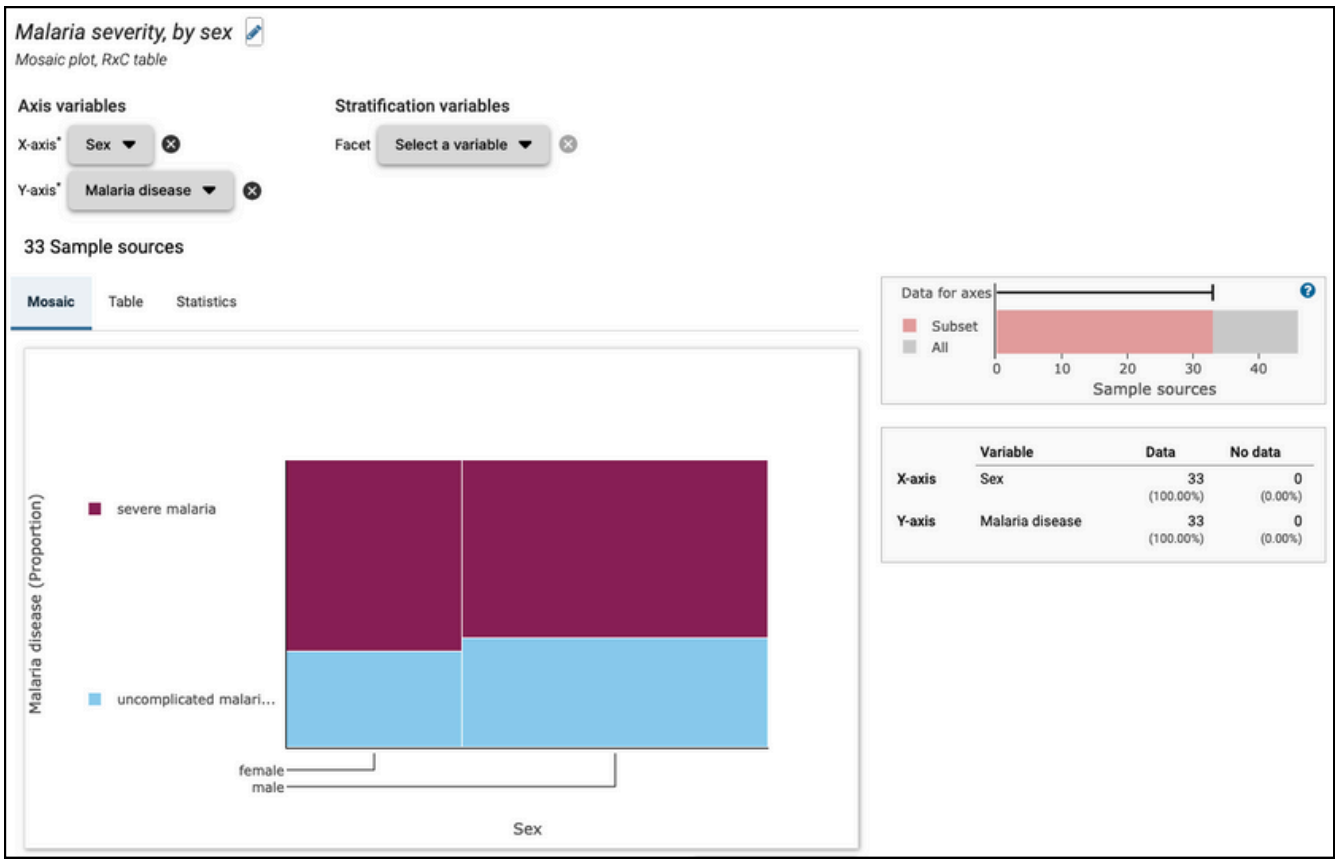


- Click the arrow next to h_P6_I9_M11 to go back to the search strategy and see the list of genes associated with this eigengene, as shown below.



2. Counts and proportions. For example, in this dataset you may be curious about whether the severity of malaria disease varies in females vs. males. To do this:

a. Click on New visualization > Counts and proportions > Mosaic plot, RxC table
b. Input X-axis= *Sex*; Y-axis= *Malaria disease*
c. Hover over the resulting graph to see the proportions of severe/uncomplicated malaria in females/males. You might observe, for instance, that males have a slightly higher proportion of severe malaria (55% vs. 53% in females).



3. **X-Y relationships**. For example, in this dataset you may be curious about whether parasite density is correlated to coma scores (A score of 5 indicates good motor response, verbal response, and eye movement, while a score under 5 is considered abnormal), and whether this relationship looks different in severe vs. uncomplicated malaria.

To do this:

a. Click on New visualization > X-Y Relationships > Scatter plot
b. Input X-axis= *Parasite density by RNA sequencing*, Y-axis= *First Blantyre coma score*
c. Choose a stratification method- Facet= *Malaria disease*
d. Examine the resulting graphs. You may observe, for instance, that many participants with severe malaria have coma scores <5 but there is no obvious correlation with parasite density.

"**Download**" tab

The download tab allows you to download the data that is represented in the study explorer and work with it on your own.

Contact **help@VEuPathDB.org** for assistance, or with questions or comments about the Study Explorer!