Host- Pathogen Interactions: WGCNA Study Explorer

What is the VEuPathDB Study Explorer? It 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

Learning objectives: Navigate the study explorer for the WGCNA dataset of metadata from dual (host and parasite) transcriptomic analysis of 46 Gambian children infected with either severe malaria or uncomplicated malaria to
- Browse metadata and use visualization tools
- Conceptualize co-expression networks and choose a host or parasite module of interest

1. Navigate to the WGCNA study explorer:
   https://hostdb.org/hostdb/app/workspace/analyses/DS_82dc5abc7f/new/learn
2. **Examine the layout of the page**- there is a diagram at the top (red arrow in image below) and a set of tabs that you can toggle between (red box below). The "Learn" tab explains the basic features of the study explorer.

3. **Examine the dataset diagram** at the top of the page.
    a. How many sample sources (i.e., study participants) were in this dataset?
    b. How many assays are represented in the dataset?

4. **Learn more about the dataset**. Click on the "**View Study Details**" tab and read the study description. How were the blood samples assayed?

5. **Learn more about the metadata (variables) in this dataset**. Click on the "Browse and Subset" tab. Examine the list of variables on the left. Click on any variable to see its distribution. Note that the distribution of **continuous variables** like *Age* is represented using a histogram, while the distribution of **categorical variables** like *Sex* is represented using a frequency table.

   The "Find a variable" search box allows you to search for variables. In this dataset, you can see variables associated with
    a. **Sample Source** (study participants), such as *Age*, *Sex*, and *Malaria disease*
        i.   What is the mean age of the participants?
        ii. What percentage of participants have severe malaria in this dataset?
    b. **Sample** (blood sample) such as *Percentage parasitemia*
        i.   How many participants had 5-10% parasitemia? (Hint: you may have to adjust the bin width for the X axis in the controls under the graph)
    c. **Assay** (WGCNA) such as host and parasite eigengenes, representing co-expression networks/clusters/modules (more on this later)

6. **Explore associations between two or more variables**. Click on the "**Visualize**" tab and click on "new visualization" to open a menu of visualization apps.
    a. Does the severity of malaria vary in females vs. males in this dataset? What app would you use to answer this question?
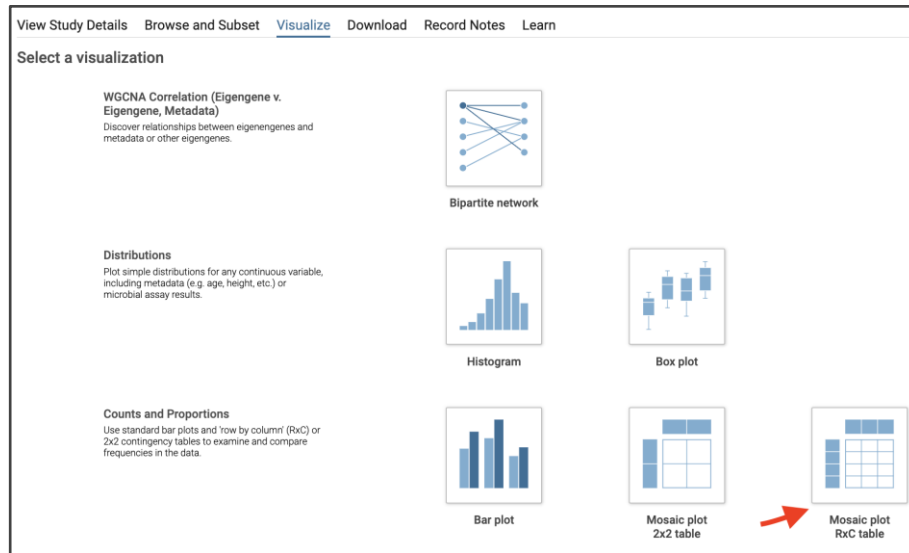
    i.   Click on New visualization > Counts and proportions > **Mosaic plot, RxC table**

    ii.   Input X-axis= Sex; Y-axis= Malaria disease

    iii.   Give the visualization a name

    iv.   Hover over the resulting graph to see the proportions of severe/uncomplicated malaria in females/males.
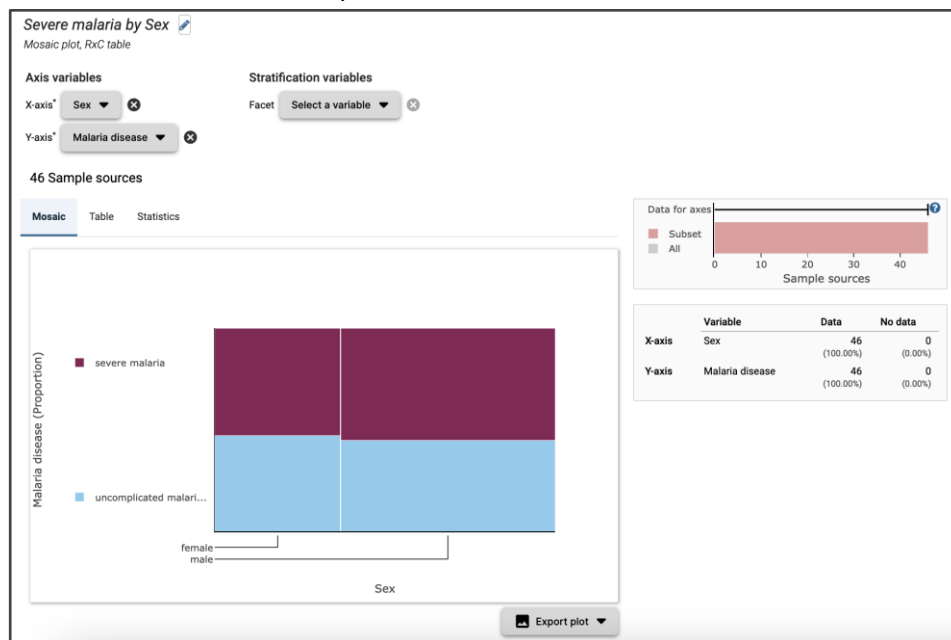
    v.   Is there a difference in the proportion of malaria between males and females?

b. Is there a difference between parasite burden (percentage parasitemia) between children with uncomplicated and severe malaria? What app would you use to answer this question?
   i. Click on Visualize> New visualization > Counts and proportions > **Box plot**
   ii. Input X-axis= Malaria disease; Y-axis= Percentage parasitemia
   iii. Is there a difference in parasite burden between the two groups?
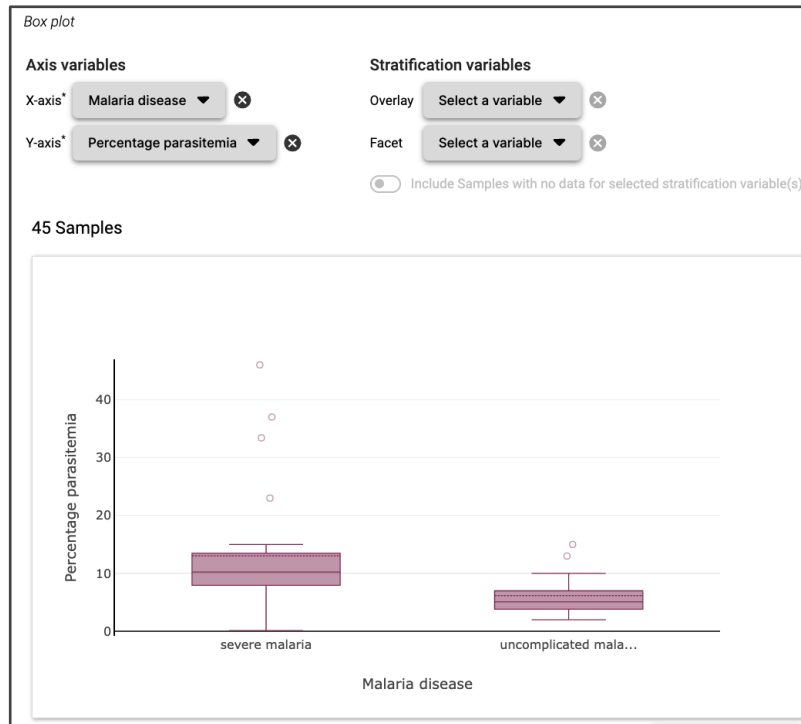
c. In this dataset, Weighted Gene Coexpression Network Analysis (WGCNA) was performed on the dual transcriptomic data to identify groups of genes that have similar expression across samples. The analysis considers the whole transcriptome data set and clusters coexpressed genes into groups called **modules**. Each module is represented by an **eigengene**, which is created in the WGCNA analysis and is an imaginary gene whose expression profile represents an average gene within the module.

Previous studies have shown a correlation between the expression levels of host genes and circulating parasitemia. In this dataset, parasite load differed between the groups with severe and uncomplicated malaria (as we visualized in step 7 of this tutorial). We were interested in identifying genes correlated with parasite density.

i. Click on New visualization > **Bipartite network**

ii. Input the Data 1, choosing Host eigengene (wgcna) from the drop-down menu. For Data 2, select continuous metadata variables

Iii. Choose a correlation method, e.g., Pearson correlation to identify linear trends

iv. Click on "Generate correlation results"

v. Visualize the correlation results. What host eigengene is positively correlated with Percentage parasitemia?
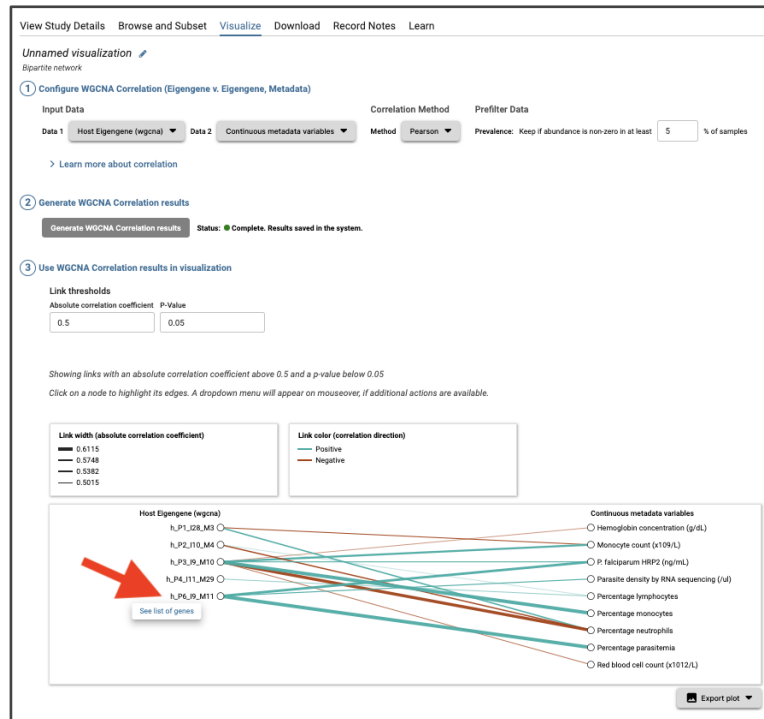
vi. Click the arrow next to h_P6_I9_M11 to go back to the search strategy and see the list of genes associated with this eigengene. Do you see any interesting genes in this list?