Visualization of Metagenomics Data

LeAnn Lindsey, Kimberly Truong, Lourdes Valdez
Final Project, CS 6630
Fall 2020

Table of Contents

Overview and Motivation

Advances in genomic sequencing technology have provided scientists with vast quantities of data to investigate scientific questions. It is now possible to obtain DNA and RNA sequence data not only for a host, but also to obtain metagenomic sequencing for all of the microbes within a specific host site, such as a mammalian gut. This metagenomic sequencing provides not only the genome sequences of the microorganisms present in the system, but also provides insight into the abundances of each species, and a phylogenetic tree of species present. It is known that the presence and abundance of specific species of microorganisms are linked to metabolic disease, autoimmune responses, pathogen detection and toxin metabolism. This metagenomic sequencing data is extremely rich but complex, and difficult to mine for information. Biologists often need to sift through this data searching for a specific gene or pathway which may be upregulated or downregulated under specific conditions and of interest in their research.

While many different tools for visualizing genomic data do exist, very few are interactive enough to be useful during the research process and are more often used to create high quality images for publication. We highlight out some of the challenges with current visualizations:

- Only a few genomic visualization tools have been extended to visualize metagenomic data, which has a higher level of complexity than genomic data.
- Do not allow a scientist to easily visualize two different experimental conditions at the same time
- Difficult to easily obtain functional pathway information from a gene
- Difficult to install and get working on various software systems, require time to learn to use and interpret output
- Many options available but most only provide one type of visualization

Our visualization project is to provide a useful tool to help scientists explore a metagenomic data set. Denise Dearing's laboratory studies microbial detoxification, and we have partnered with her lab to visualize a specific data set acquired last summer by Rodolfo Martinez-Mota, which explores microbial cardenolide detoxification in wild black-eared mice. Monarch butterflies have evolved a resistance to plant toxins, specifically, they feed on milkweed which has high levels of cardenolides. Monarch butterflies migrate each year to overwintering sites in the southern United States and Mexico and during this migration season, predators such as the wild black-eared mouse feed on the butterflies. The goals of the study are to determine the role of the gut microbiome of black-eared mice on detoxification of the cardenolides ingested with a monarch based diet. The Dearing lab has investigated the role of the microbiome using 16S rRNA gene marker sequencing, which provides some information about the species abundances and genes of interest, but they have not yet fully processed and analyzed the metagenomic data set acquired in the study.

Data

Our metagenomic shotgun sequencing data consists of 14 samples of wild black-eared mice under the following experimental conditions:

- Monarch/Wild 6 samples
- Non-Monarch/Wild 2 samples
- Monarch/Experiment 2 samples
- Non-Monarch/Experiment 4 samples

Samples labeled “Monarch/Wild” were collected in the wild during Monarch season and are assumed to have eaten Monarch butterflies as a part of their normal foraging.

Samples labeled “Non-Monarch/Wild” were collected during the Non-Monarch season and are assumed to have not eaten Monarch butterflies during that time period.

Samples labeled “Monarch/Experiment” were taken from mice that were captured during Monarch season and fed a diet of 5 Monarch butterflies per day for a period of 48 hours.

Samples labeled “Non-Monarch/Experiment” were captured during Non-Monarch season and taken into captivity for the same period of 48 hrs but were not fed Monarch butterflies.

Every sample contains gene family abundances, functional pathway abundances, and taxonomic abundances (quantitative data).

Questions

**What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?*

With this data set, we aimed to address the following questions:

- What is the role of the black-eared mice gut microbiome in cardenolide detoxification?
- What are the differences in the gut microbiome of black-eared mice on a Monarch butterfly diet vs on a non-Monarch butterfly diet?
- Are any genes up- or down-regulated in mice with Monarch vs non-Monarch diets?

Over the course of our design process and data analysis, our questions became more specific to the species level across samples: Why do some samples seem to have higher abundance in many species and other samples have none of those species?

Tasks

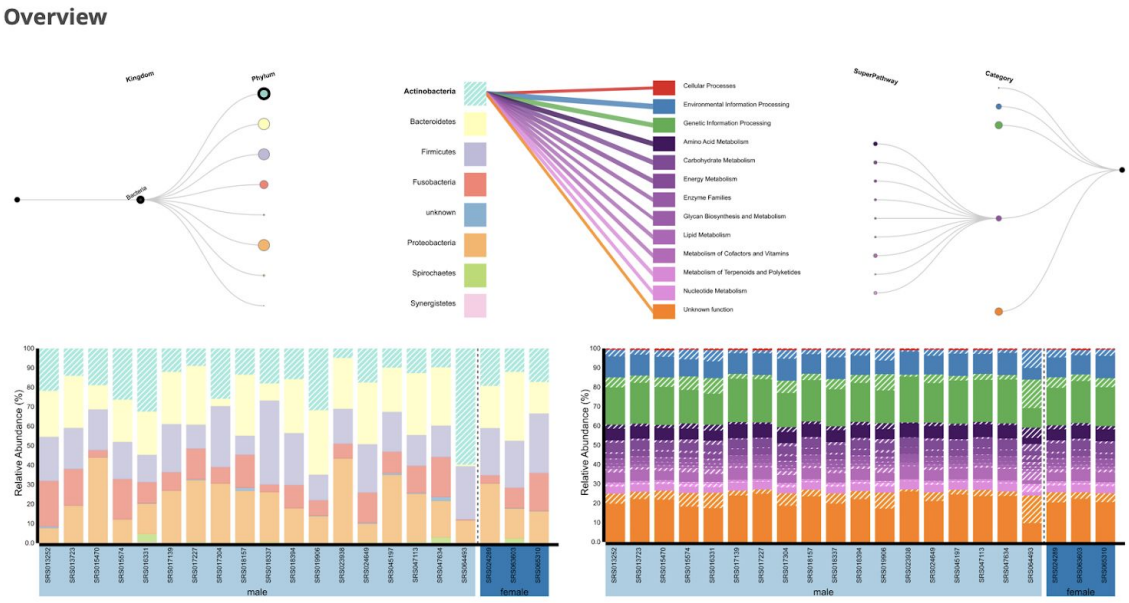
Analysis Tasks

1. Filter data by level from family to species.
2. To identify species with genes and functional pathways present in that species.

3. To compare samples and experimental groups.
4. To compare gene expression and functional pathways in different experimental groups.
5. To identify outliers and trends.
6. To discover and present interesting aspects of this specific Monarch dataset.

Domain (Users)
Our domain is microbiology, and the target audience is biologists and researchers interested in exploring this specific Monarch dataset.

Related Work



We were inspired by a similar metagenomic visualization tool named BURRITO after discovering it in a lab meeting with Denise Dearing. This later on gave us the idea to color the nodes of our tidy tree to match the data in the stacked bar charts.

Data Processing

Taxonomy data: The raw data was obtained from MetaPhlan2 (with a lower threshold for rare species) and then reformatted to CSV in Unix. The tabular data contain 136 rows representing unique taxonomic units. We also had to shorten the sample names and append the experimental condition to the name. The csv data was transformed into a hierarchy to create a Tidy Tree using d3.stratify.

Metagenomic data: Metagenomic sequence data was processed with Humann3 software from the Huttenhower Lab at Harvard University, to obtain gene family abundances and functional

pathway abundances (.tsv). It is too large to visualize, so we had to perform feature reduction using the Singular Value Decomposition, a numerical linear algebra technique which identifies the most important components in the data. At first, we reduced the data from 280,000 rows to 10,000 rows using features of interest. We realized 10,000 was still too large to render so we further reduced it down to 1000 rows.

Data Cleaning for Stacked Barchart: The stacked bar chart required the metagenomic data file to be transposed. We ended up transposing this data outside of JavaScript so the stacked barchart had its own csv file we would pass in.

Data Cleaning for Heatmap:

The two files that will be used in the heatmap are:

Combined_genefamilies_stratified.tsv	Number of Genes 280321
Combined_pathcoverage_stratified.tsv	Number of Gene Pathways 841

The genefamilies file is too large to show in the heatmap, which after some trial and error, we decided to limit to 1000 rows. We needed to make some decisions on how to filter the data and select which genes we wanted to include in the heatmap. We considered several options for limiting the data.

1. Remove all “unclassified” genes.
2. Remove all genes that are not present in a specific threshold of the samples.
3. Use Singular Value Decomposition to determine the most important features in the data
4. Using EdgeR to identify the most differentially expressed genes

We discussed the options to select the data and these were the advantages and disadvantages of each one.

1. Remove all “unclassified” genes
   Advantages: Simple
   Disadvantages: Unclassifiied and low abundance species may be the species of interest to the researcher
2. Remove all genes that are not present in at least ½ of the samples
   Advantages: Easy to implement
   Disadvantages: This option will favor genes that are present in every sample, which may not be the best way to see differentially expressed genes, which are of most interest to researchers.
3. Use Singular Value Decomposition to determine the most important features in the data
   Advantages: Easy to implement
   Disadvantages: Did not end up reducing the dataset enough and did not contain enough variation between the two conditions.
4. Using EdgeR to identify the most differentially expressed genes.

Advantages: Most informative for researchers and familiar to them
Disadvantages: More difficult and time consuming to to implement

The gene pathway file did not need to be reduced in size to be shown in the heatmap.

The format of both files had to be changed so that d3 could properly bind the data to end up with the correct number of rectangles. The format had to be changed from a matrix format with genes on the rows and samples on the columns, to a long format that has one row for each rectangle in the heatmap:

Column, Row, Value
Gene1,Sample1,Value
Gene1,Sample2,Value
Gene2,Sample1,Value
Gene2,Sample2,Value

We used python to make this change (see python notebook in github for the code).

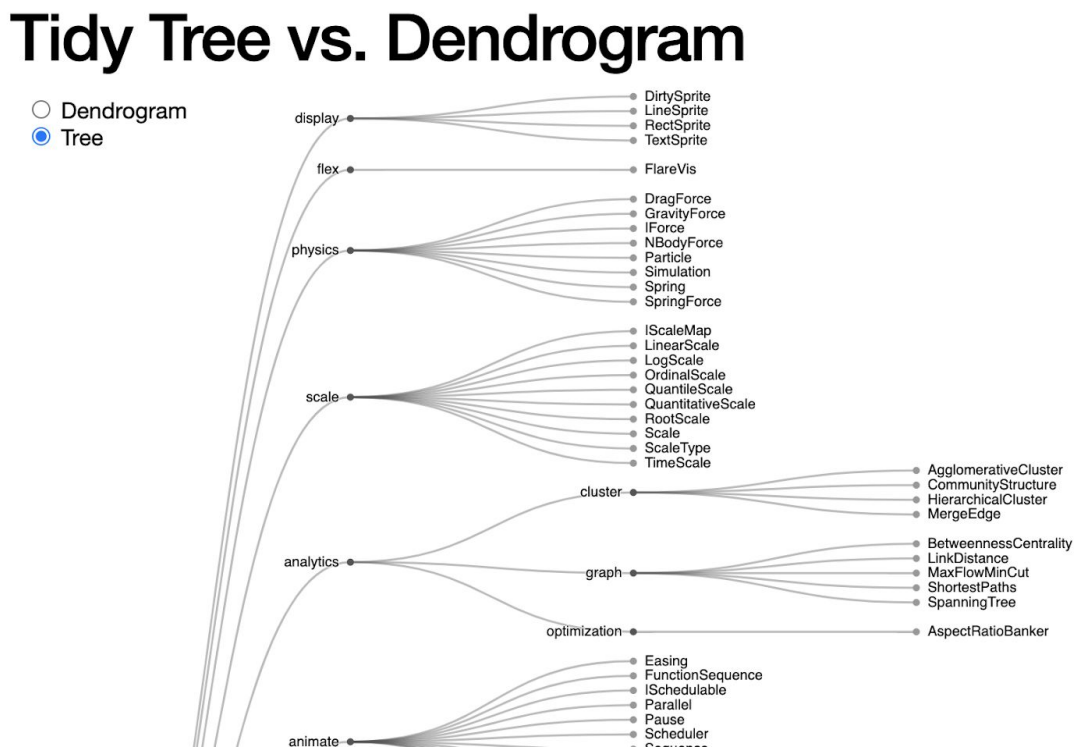
We had to modify the geneFamilies and pathwayAbundance data files to a flat format so that there would be one row per rectangle in the heatmap. We wrote a python script to do that transformation. The new format of the data is:

GeneFamily,Sample,Value,Condition

Where condition is [‘Monarch’, ‘No-Monarch’]

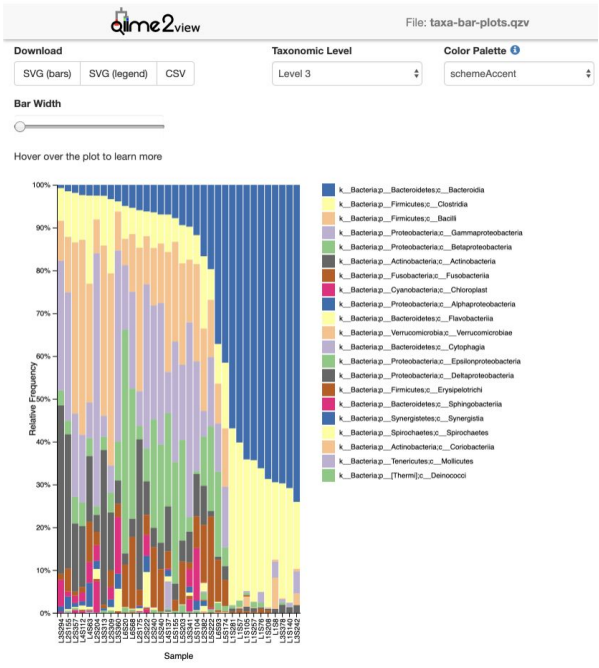
Exploratory Data Analysis

We looked at several visualizations to model our components after:



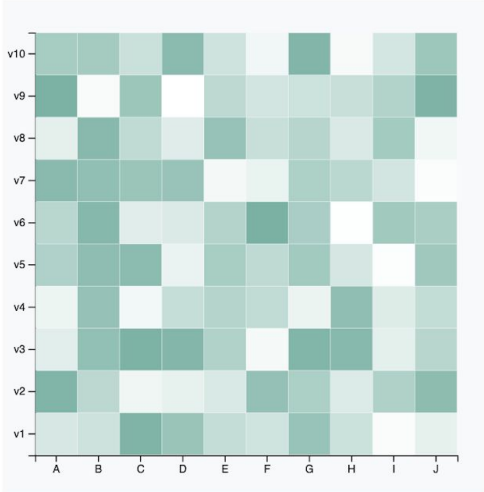
We first looked at examples of trees such as the Tidy Tree and Dendrogram example by Mike Bostock. The Tidy Tree made more sense to represent the phylogeny.

<u>Stacked Bar Chart from Qiime2</u>:



The Qiime2 website is what metagenomic researchers commonly use to look at abundances at various levels of the phylogenetic tree. We thought stacked bar charts are the most effective way to encode parts of a whole, and we didn't want to stray away from what researchers were already intimately familiar with.

Heatmap from d3-graph-gallery:



Steps:

- The Html part of the code just creates a `div` that will be modified by d3 later on.
- The first part of the javascript code set a `svg` area. It

```
<!DOCTYPE html>
<meta charset="utf-8">

<!-- Load d3.js -->
<script src="https://d3js.org/d3.v4.js"></script>

<!-- Create a div where the graph will take place -->
<div id="my_dataviz"></div>
```

```
<script>

// set the dimensions and margins of the graph
var margin = {top: 30, right: 30, bottom: 30, left: 30},
  width = 450 - margin.left - margin.right,
  height = 450 - margin.top - margin.bottom;

// append the svg object to the body of the page
var svg = d3.select("#my_dataviz")
.append("svg")
  .attr("width", width + margin.left + margin.right)
  .attr("height", height + margin.top + margin.bottom)
.append("g")
  .attr("transform",
        "translate(" + margin.left + "," + margin.top + ")");

// Labels of row and columns
var myGroups = ["A", "B", "C", "D", "E", "F", "G", "H", "I", "J"]
var myVars = ["v1", "v2", "v3", "v4", "v5", "v6", "v7", "v8", "v9", "v10"]
```

We liked this basic heatmap and adapted the D3 and Javascript code for our heatmap of gene expression and pathways.

Design Evolution

**What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?*

From our exploratory data analysis, we concluded that 4 main view objects were necessary:

(1) Tidy Tree for the phylogenetic tree;
(2) Stacked bar chart for taxonomic abundances;
(3) Heatmap for gene expression and pathways;
(4) Violin plot to compare gene expression for Monarch/No-Monarch groups.

For our proposal, we did brainstorm and consider other options for the four views:

- Zoomable Sunbursts as an alternative to the Tidy Tree.
- Zoomable circle packing and Treemaps to connect hierarchy with abundances.
- Difference charts for differential gene expression between samples.

But we thought our four original views were intuitive, effective, and easy to start with for development purposes. Below are design sheets laying out our brainstorming as a team.

Initial Design

Brainstorm Metagenomic Data Visualization

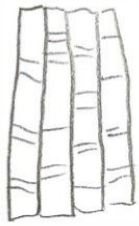
TITLE: Visualization of Metagenomic Data

AUTHORS: LeAnn Lindsey, Kimberly Truong, Lourdes Valdez

DATE: 10/30/2020 **SHEET:** 1

interactive

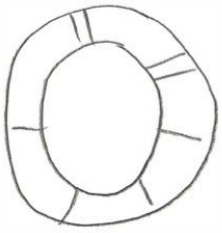
- Stacked bar chart -



Samples →

1. shows abundances in each sample
2. samples grouped by condition
3. click to drill down into Family → Phylum → Class → Species

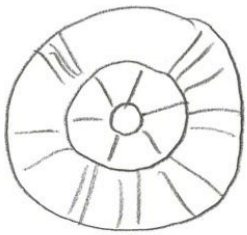
- donut chart -



clickable to drill down but - difficult to see samples compared

in general - difficult to measure/compare circles

- Zoomable Sunburst -



Can show more layers of hierarchy - but similar problems to donut chart

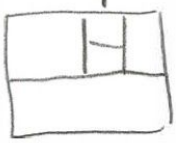
- tidy tree/dendogram -

1. traditional way to show taxonomy

- heatmap -

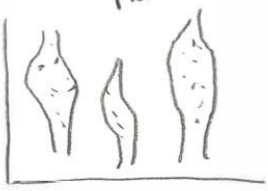
familiar, add interactivity to show gene/functional pathway on hover

- treemap -



could show abundances

- Violin plot -



Violin plot shows differentially expressed genes — with advantage of showing sample distribution and sample variation

- difference chart -

genes

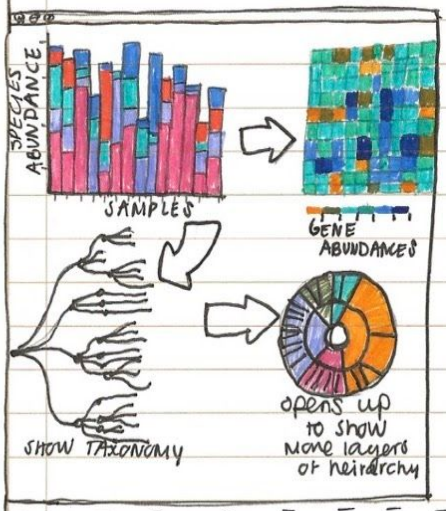
could show differentially expressed genes

- Zoomable Circle Packing -

Could show hierarchy & abundances but not familiar to read & difficult to see granularity

SUBJECT:

LAYOUT:



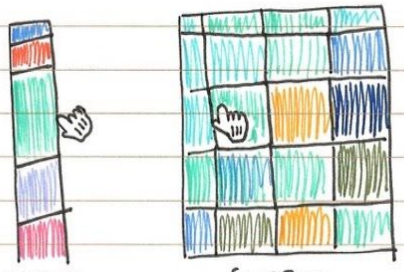
TITLE: visualization of Metagenomic Data

AUTHORS: LeAnn Lindsey, Kimberly Truong, Lourdes Valdez

DATE: 10/30/20 SHEET: 3

TASK: Determine role of gut microbiome of black-eared mice on detox of cardenolides.

OPERATIONS:

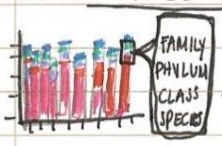


SNIPPET FROM STACKED BAR

SNIPPET FROM HEATMAP

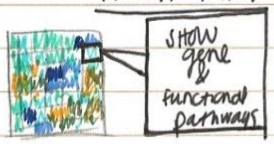
FOCUS/ZOOM:

STACKED BAR CHART



ON CLICK

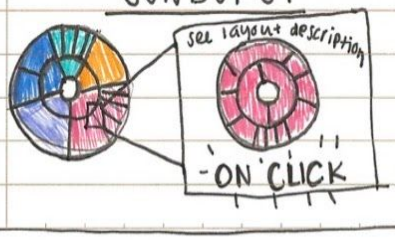
HEAT MAP



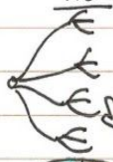
ON HOVER

PHYLOGENETIC TREE

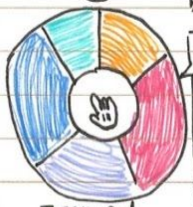
SUNBURST



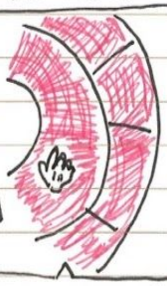
UNEXPANDED VIEW



User will click on branch of interest for zoomed-in view.



zoomed in VIEW FROM SUNBURST



CLICKING ON A REGION LETS YOU EXPLORE IT DEEPER

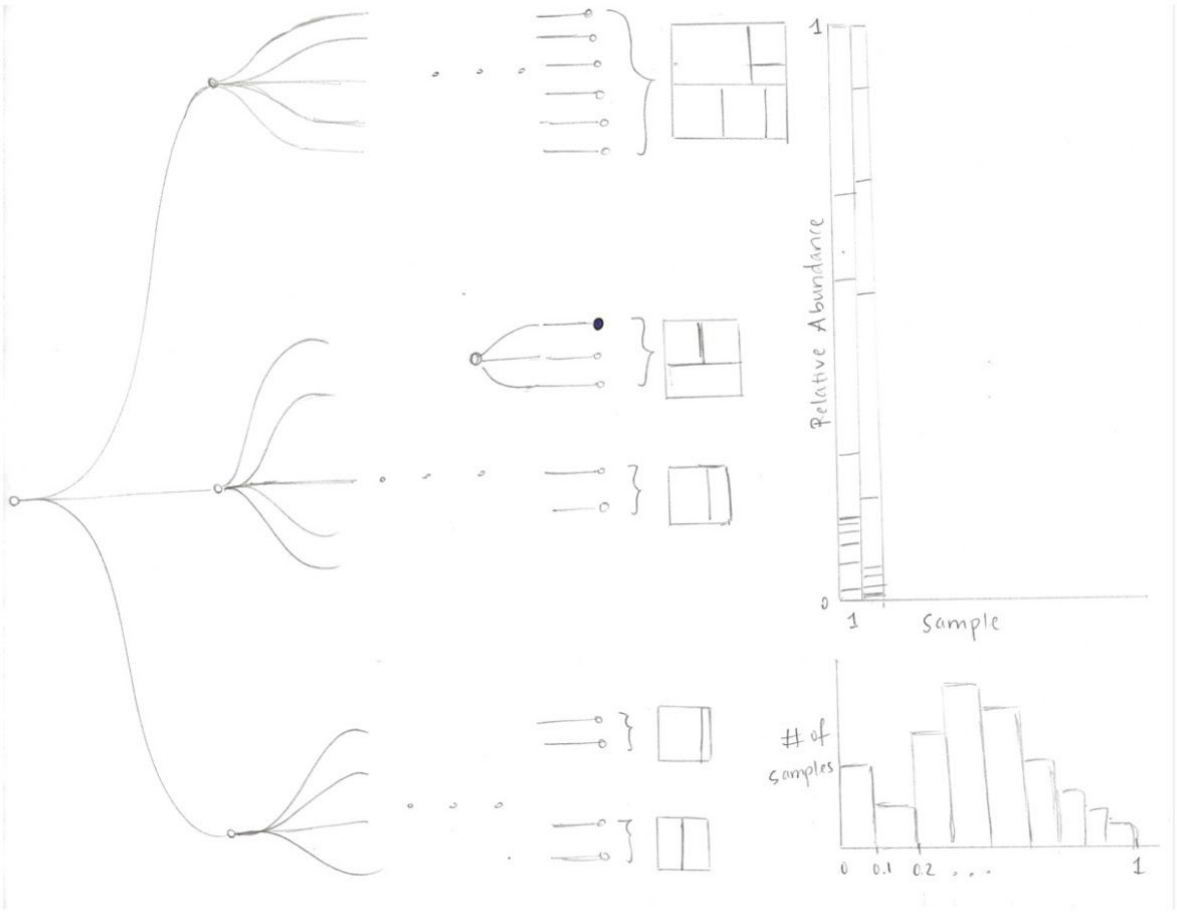
DISCUSSION:

PROS:

- Allows story to be told.
- Good use of focus allows us to give details of data w/out cluttering the webpage.

CONS

- Too much interactivity or too much data can break the page.
- Might need to simplify some visualizations.

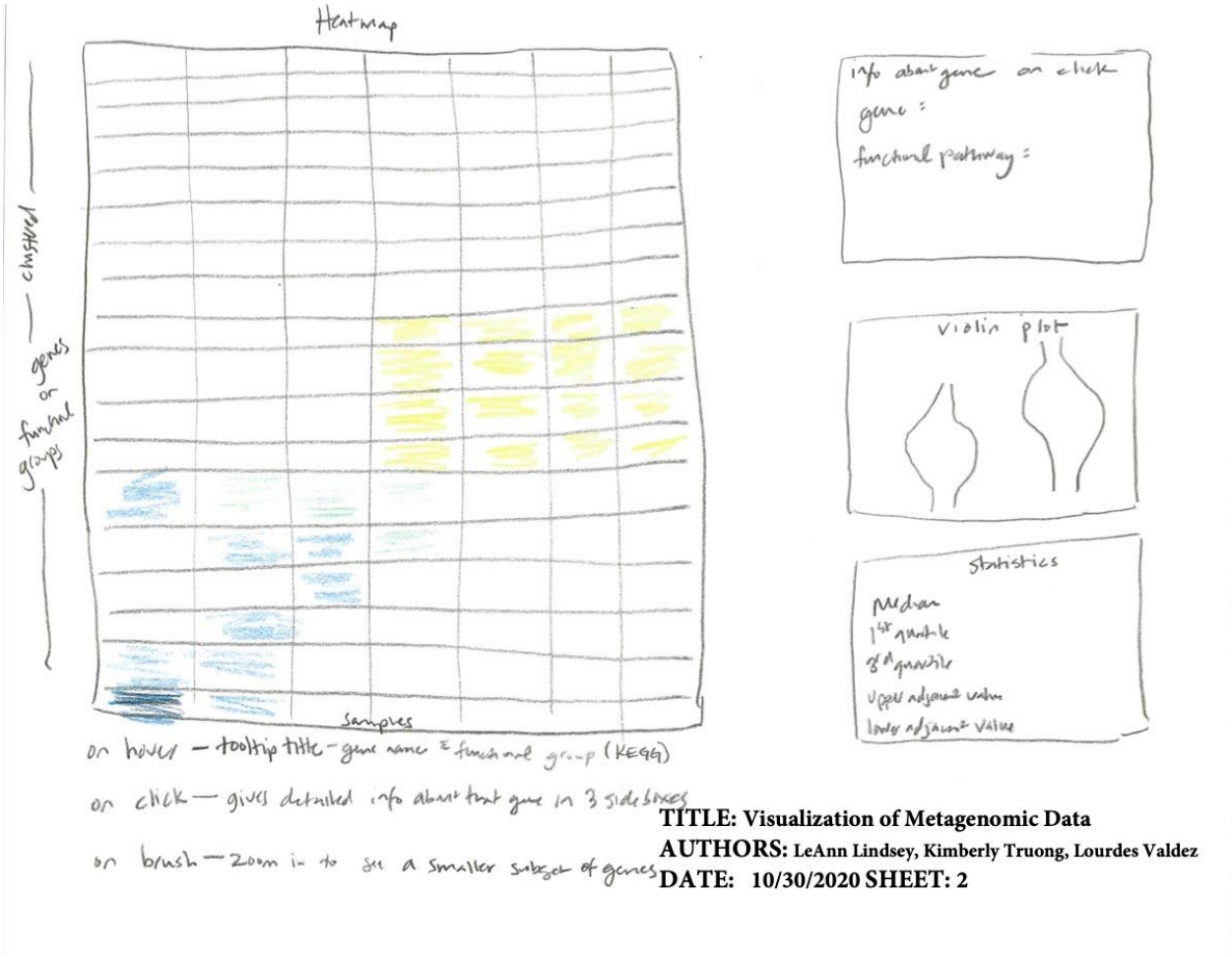


TITLE: Visualization of Metagenomic Data
AUTHORS: LeAnn Lindsey, Kimberly Truong, Lourdes Valdez
DATE: 10/30/2020 SHEET: 4

Together, as a team, we decided that for the first part of the webpage, the user can view taxonomic abundances by interacting with the tree and eventually drill down to the species level. We also included a histogram of per-sample abundances for a selected species on click as a

way to show all our data. We thought this would give the user some perspective at first glance if the data reflected 10 vs 100 total samples. We had this in mind when we had planned to visualize a larger number of samples. But after we chose a dataset of 14 samples, we scrapped the histogram as it was no longer meaningful.

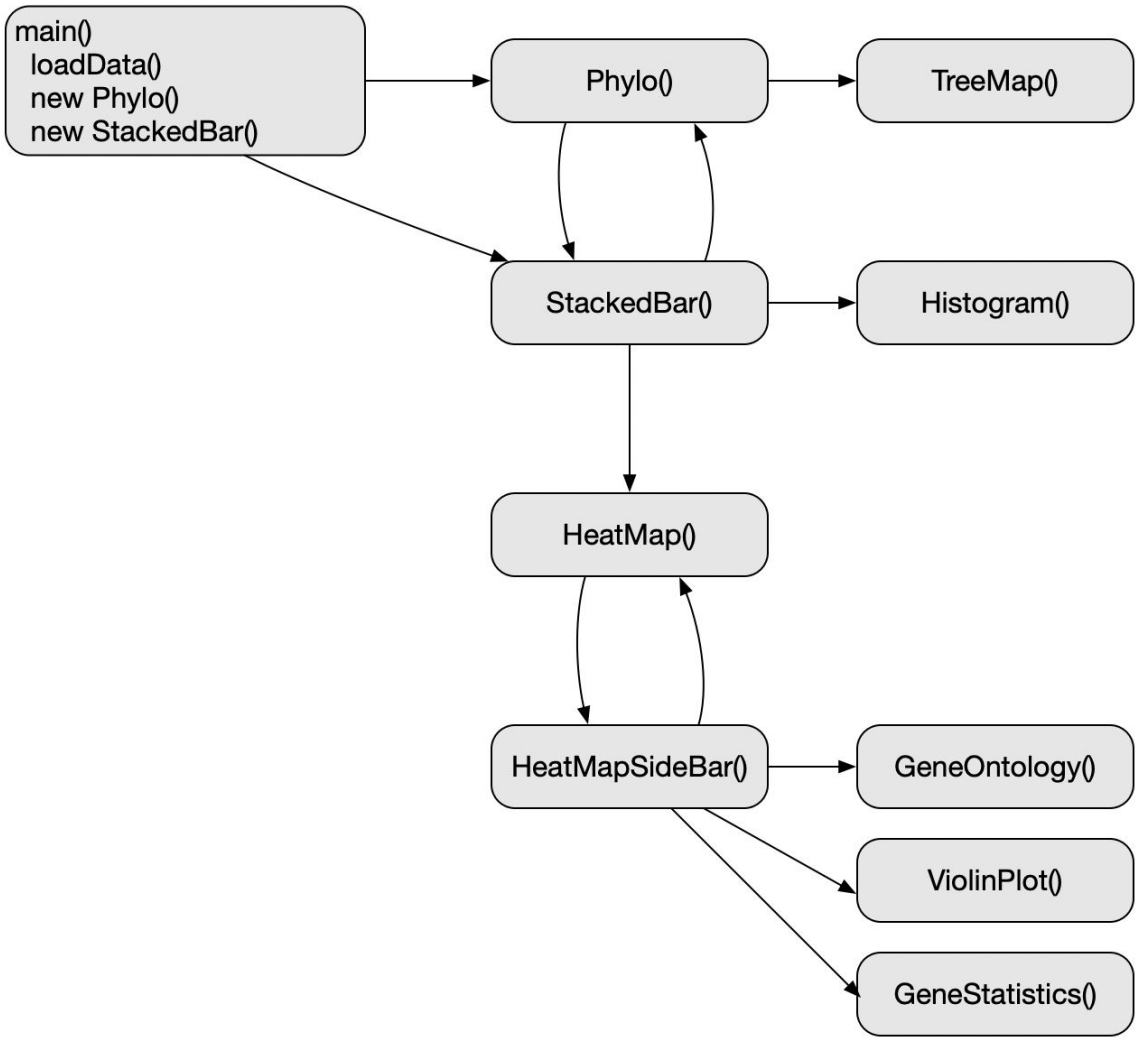
At the leaves, treemaps will appear to show the average abundances of related species within a genus.



Scrolling down, the user would find the heat map and violin plots underneath with text about a selected gene on click and statistics of gene expression across the two experimental conditions.

To get started, each component was implemented independently by a person in the group.

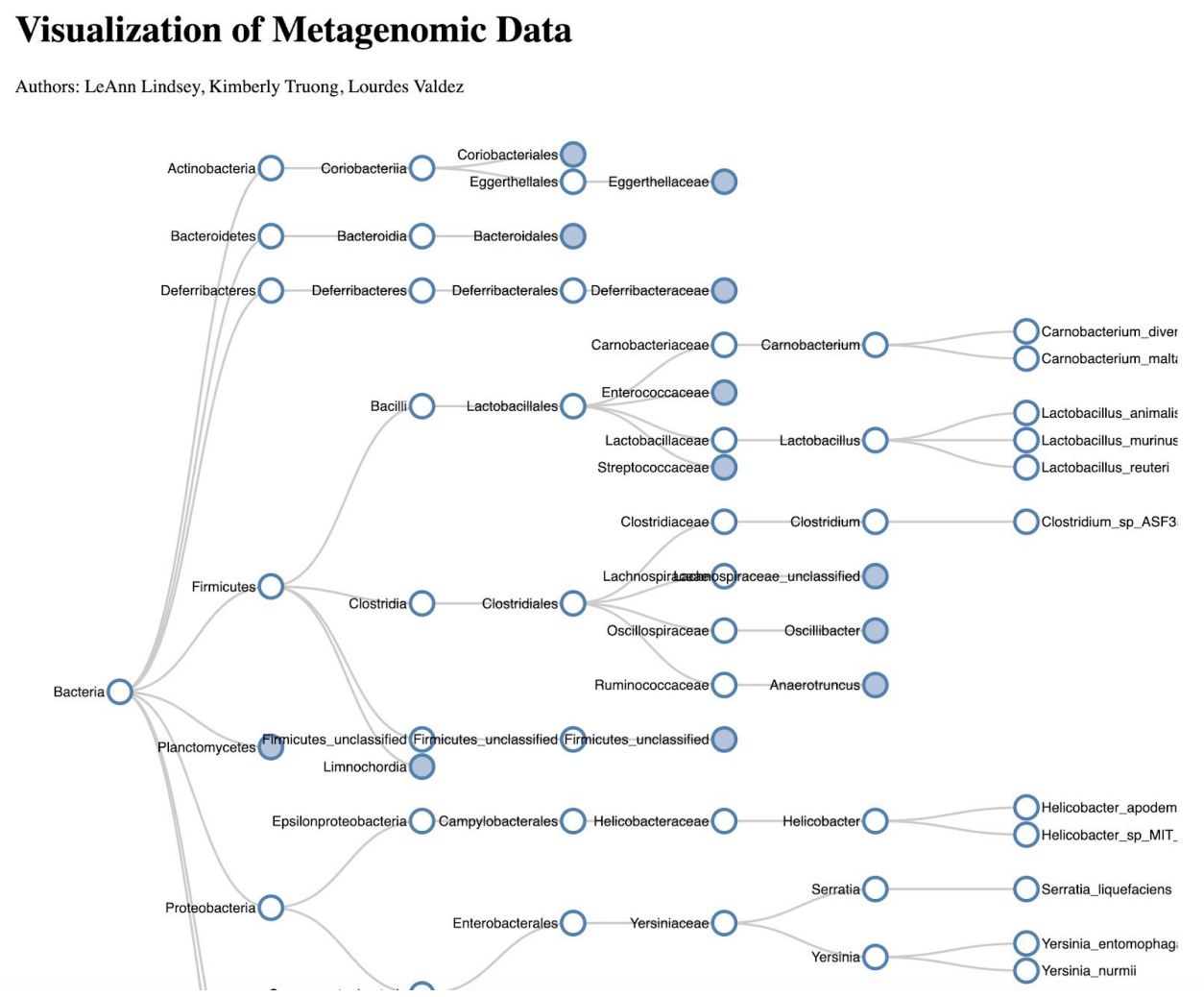
Component Integration Design



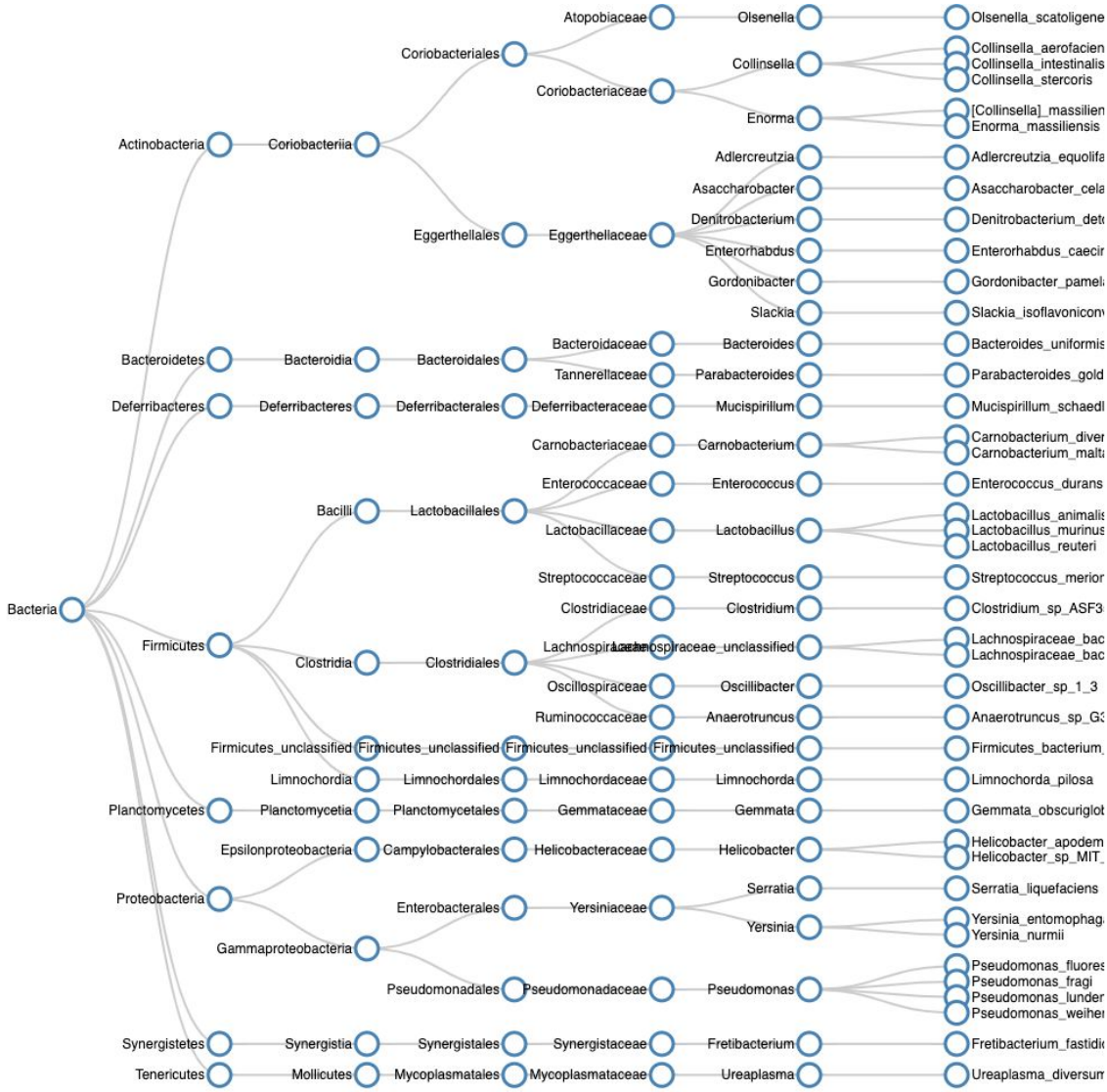
Tree

We first visualized the entire tree on the webpage. To our surprise, the full tree was displayed and the user could see every part of the tree clearly. Despite the nice resolution, the height of the tree was far too big at 2400px.

Next, we implemented an interactive version of the tree in which all the nodes can expand or collapse their children upon click. Nodes with hidden children were encoded with a blue fill, and nodes with their children expanded are encoded with a white fill color.



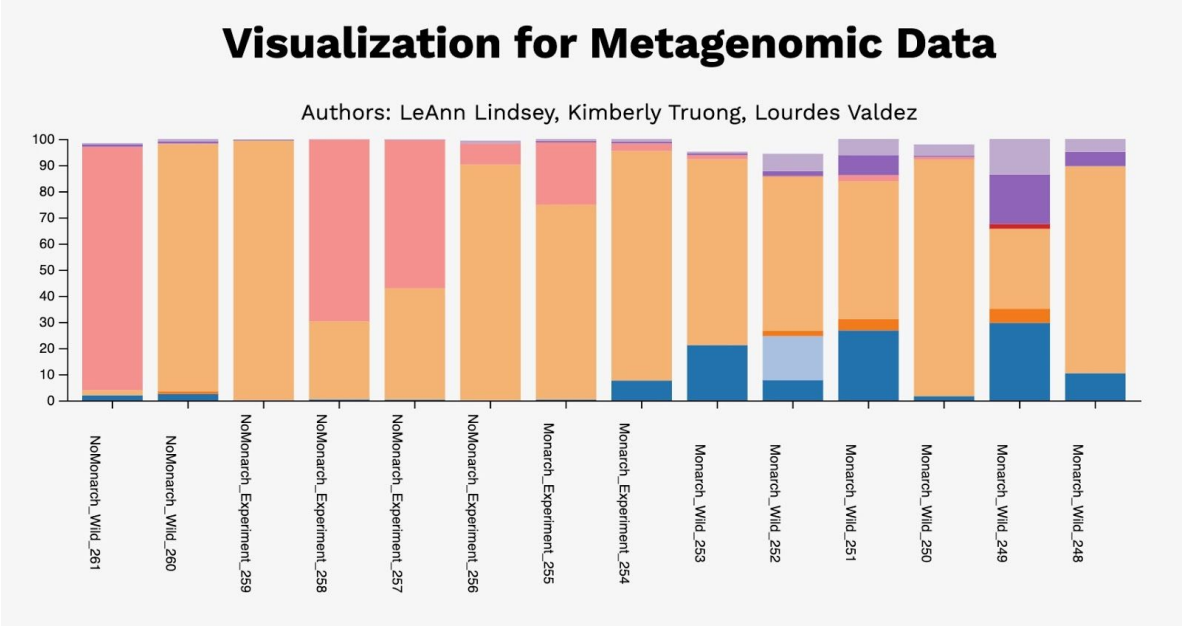
A view of the entire tree after expanding all the nodes:



We thought this still looked squished and the species' names were being cut off. Many species' names are redundant--they follow the format of genus_species-name. We considered chopping off the genus name but thought it was weird to have species with names like "sp_1_2". Thus, we maintained the original annotations.

Stacked Bar Chart

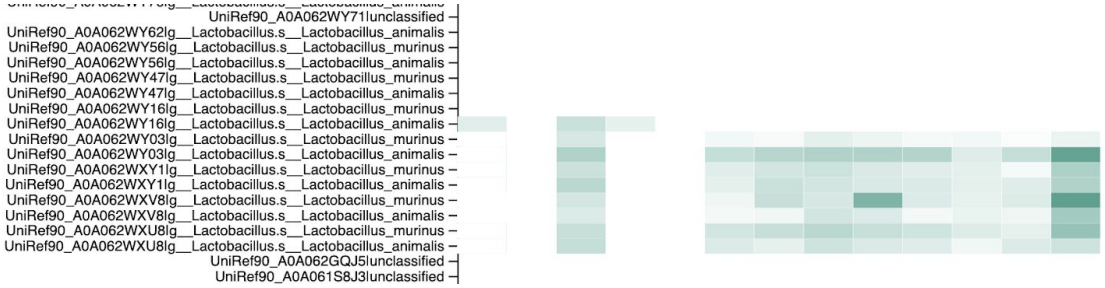
This was our first successful implementation of the barchart:



The original data was transposed (rows changed to columns and columns changed to rows) in order to be able to draw the stacked bar chart correctly. Once the stacked bar chart was drawing, we decided to increase the height of the chart so that the species with smaller abundances were more clearly visible.

Heatmap

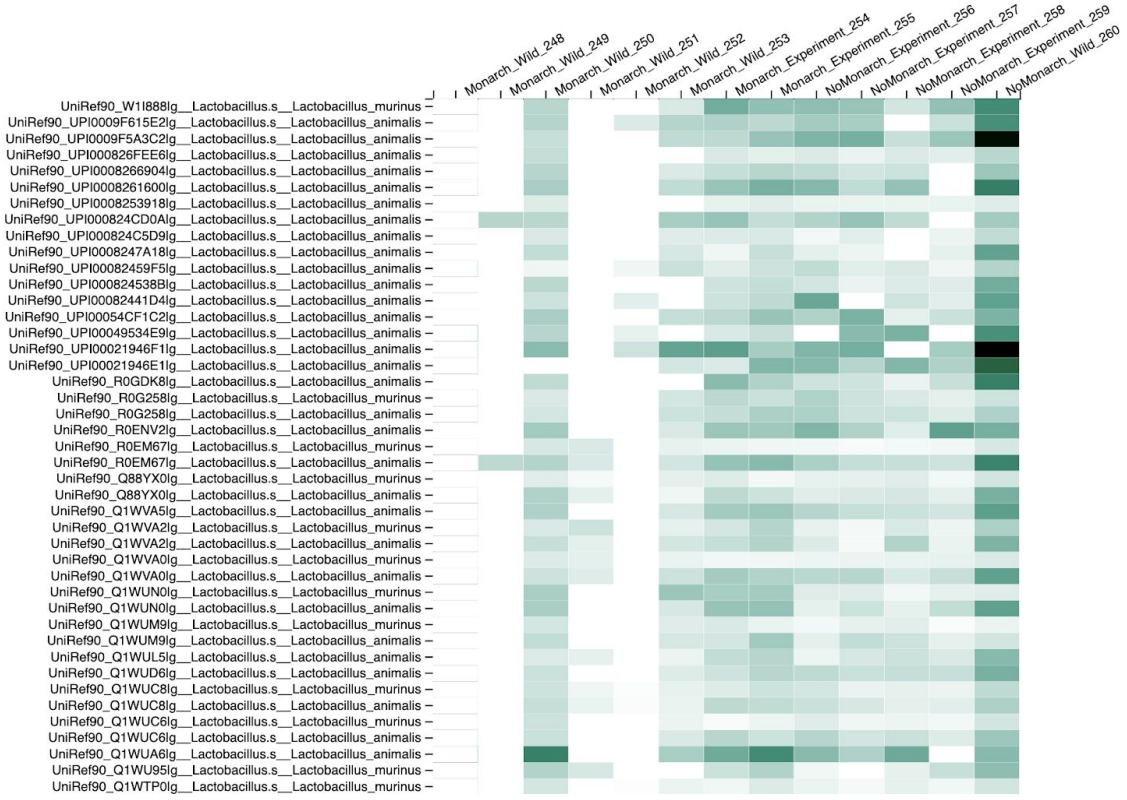
Heatmap v1



Notes on Heatmap v1:

This was the colormap used in the sample, and we decided to change from our original idea of having a divergent colormap to a one color heatmap. It made sense to have white be the zero value, since there are no negative values in the data matrix. We used a small random subset of data for the initial prototypes. The final version will use data that is filtered by differential expression analysis. We created a threshold function in python to subset rows having a minimum threshold value and a minimum threshold count. The code is in the Jupyter Notebook in the github. We thought a long heatmap that scrolled off the page would not be usable, but length is not really a problem. We scaled the size of the box to the length of the file so that if a larger file was inputted, it would not change the size of the heatmap boxes.

Heatmap v2

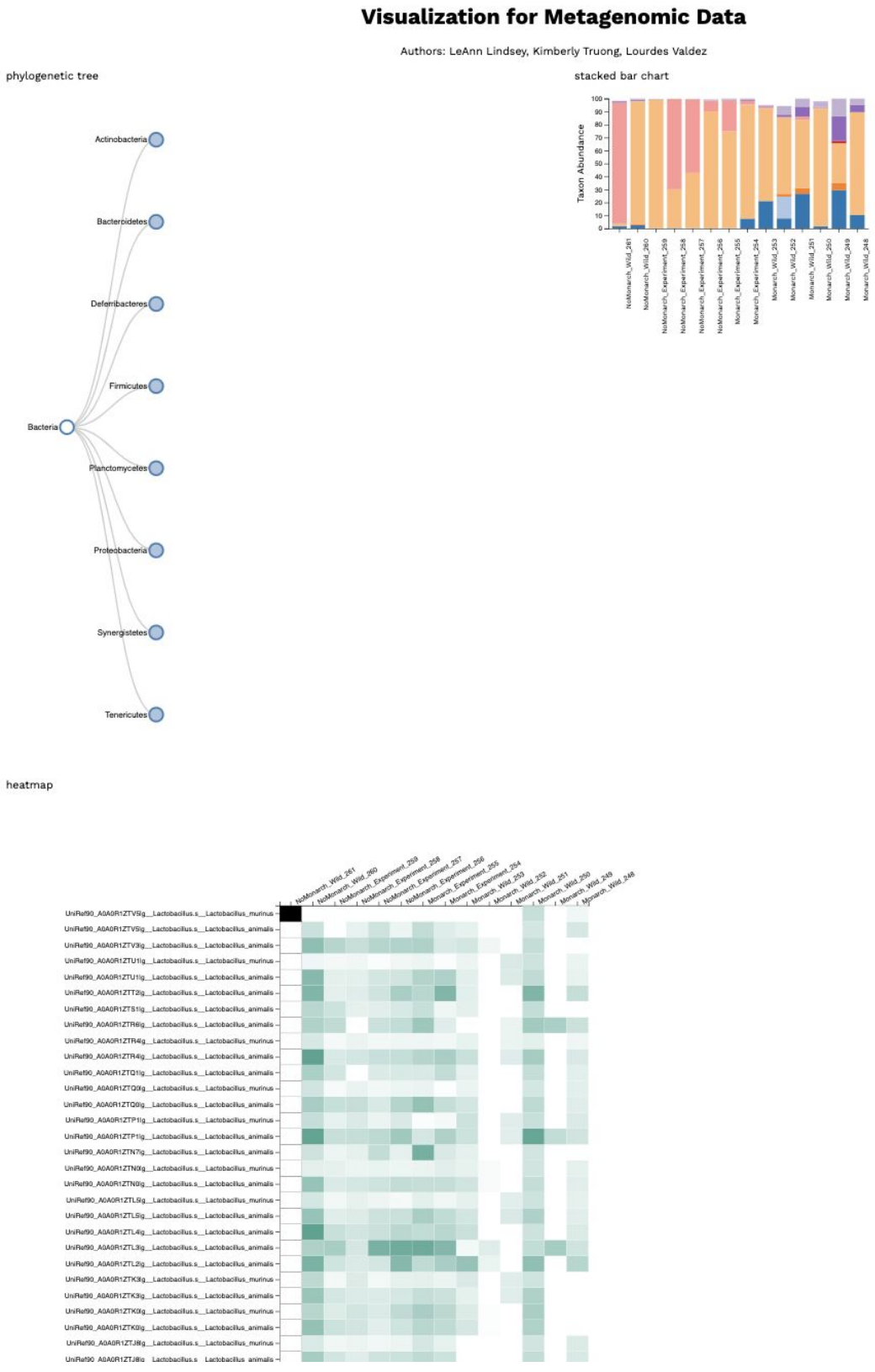


Notes on Heatmap v2:

At this phase, we noticed that there were some samples that looked very different than other samples. It could be an error in the sample or quality of the sample, or some biological factor we are not accounting for. It was interesting to notice that we could already learn about our data even from looking at a random subset of data.

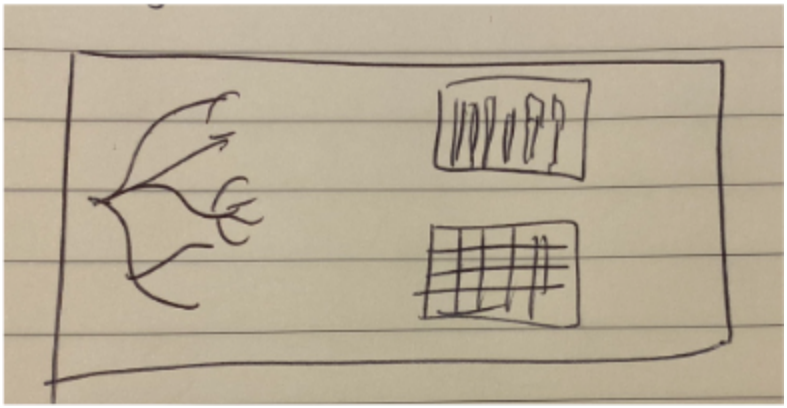
Integration of Components

Here is a picture of what the webpage looked like after we integrated our individual components:



Project Milestone

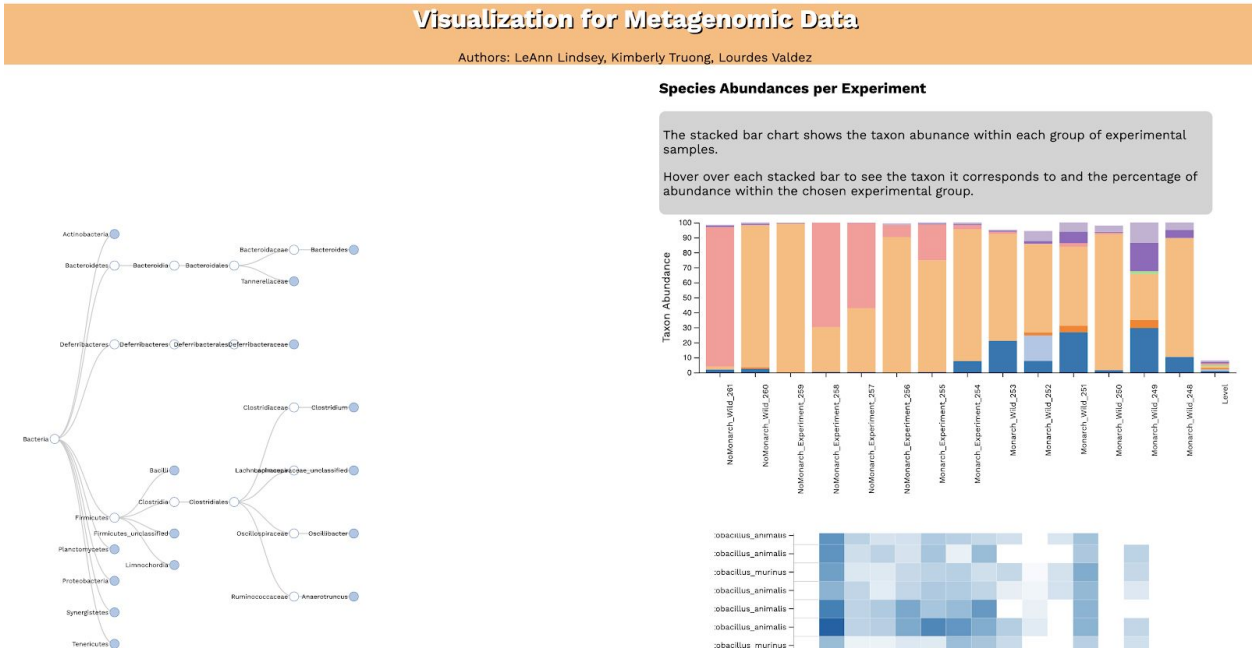
Our meeting with our project TA, Youjia, brought up the concern of how all three charts would connect with each other. In particular, the heatmap at the bottom might be difficult for users to navigate and look at the two components. Youjia suggested putting the heatmap at the bottom right like so:



There were several improvements following the milestone meeting:

- Added bottom and right scroll bars and a zoom in-and-out feature for the tree
- Added bottom and right scroll bars for the heat map

With these improvements and Youjia's suggested layout, the visualization looked like this:



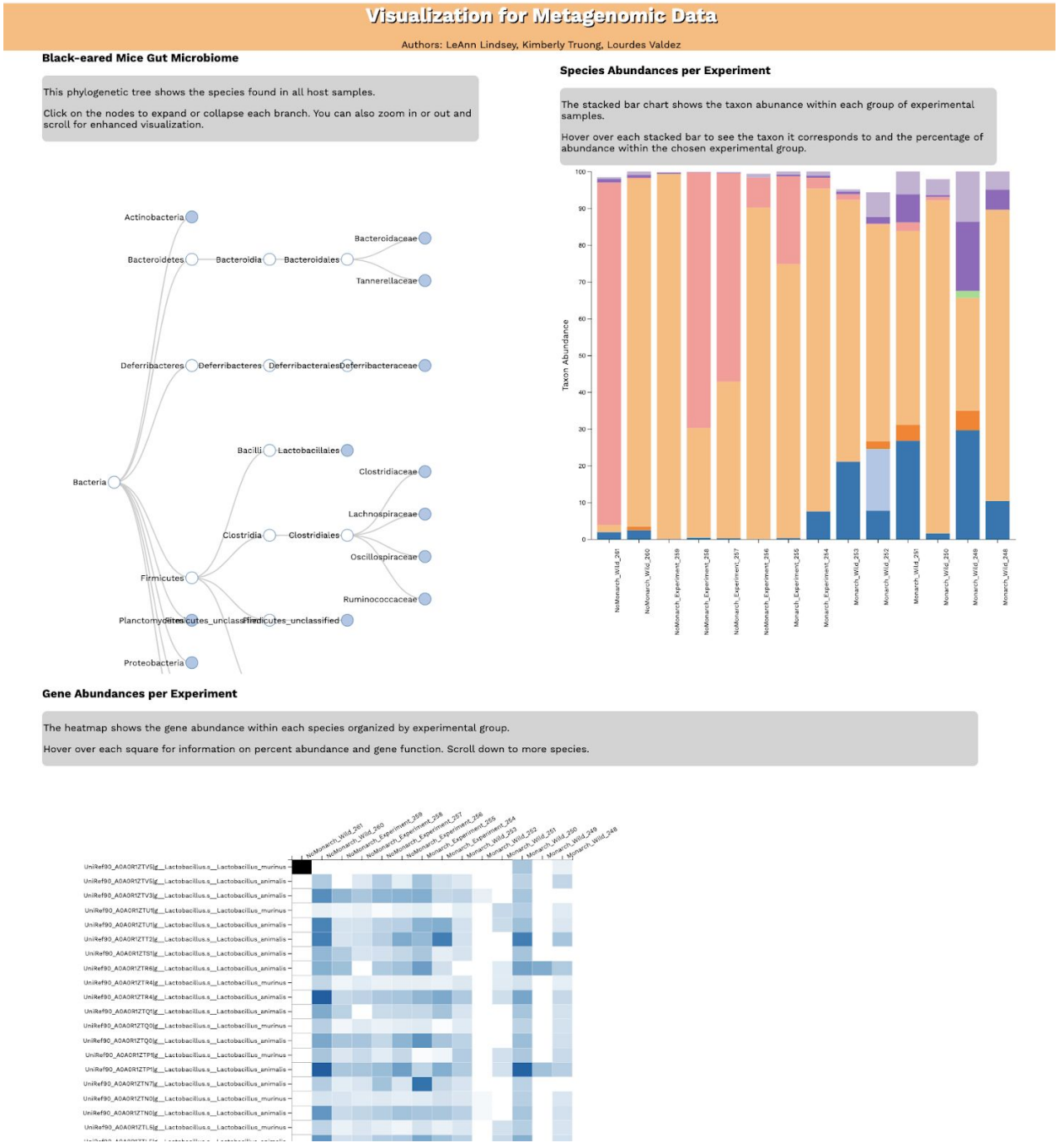
We found this change made the heatmap look very cramped and even with scrolling, it is hard to see a decent swath of genes at once. The zoom feature was also sometimes hard to control

and made the tree fly off the page. We implemented it, but we ended up just using a scrollbar instead.

Integration v2

We decided it was best to stick to our original layout in the proposal. The user has the option of two modes of exploration here:

(1) Exploring abundances in the stacked bar chart based on levels of the tree.
(2) Looking at gene abundances across the two experimental conditions with the heat map and violin plot.

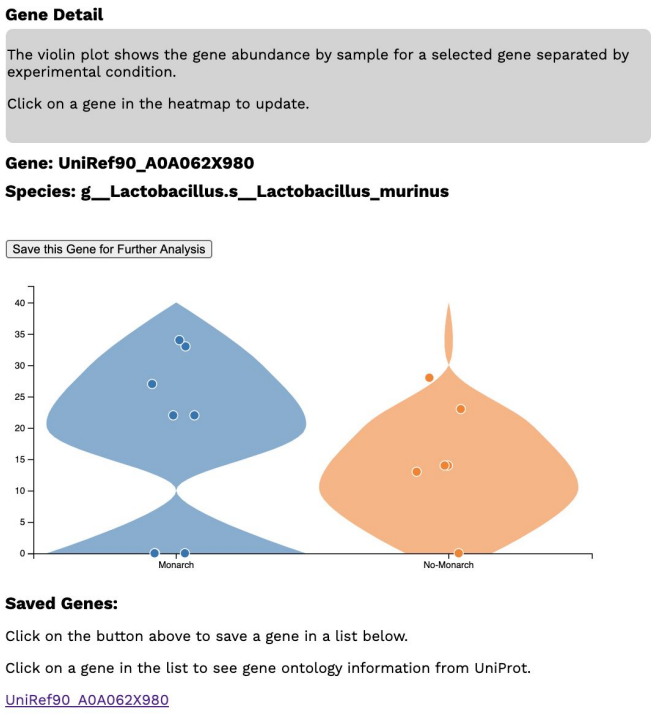


Interactivity & Further Improvements

We spent the rest of our design process, adding more interactivity and elements to coordinate these views such as:

- Mousing over an expand-able node displays a tooltip with its number of children
- Mousing over a cell on the heatmap displays a tooltip with gene expression data
- Mousing over a bar in the stacked bar chart highlights that species in all samples
- Mousing over the legend greys out that species, and clicking on a species in the legend removes that species from the stacked bar chart.
- Highlighting the x-axis labels makes them larger and easier to see, sample by sample
- Expanding the level of the tree updates the bar chart abundances according to the level of classification
- Clicking on the x-axis label brings up the appropriate sunburst plot at the level that you are looking at in the tree.
- Added info panels to describe each view with instructions on how user can interact with each view
- Integrated the color schemes with a universal mapping of the color data to a specific species
- Colored nodes on the tree only at the level that the bar chart is showing

We also added two new views of the data--a violin plot and a sunburst plot--for which we discuss the design process of below.



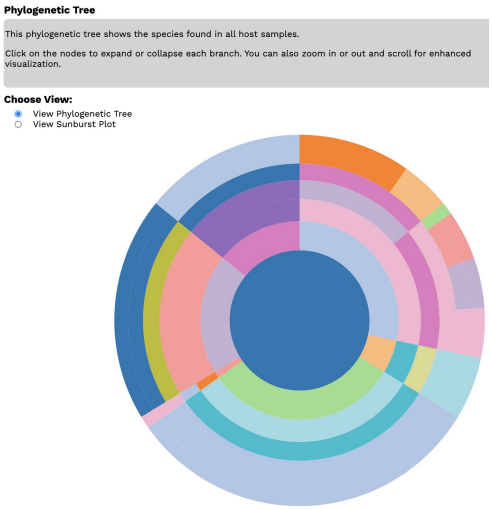
Violin Plot

The Violin plot is not as useful as we had hoped, because this dataset is very small, but it would be useful if there were 30 or more samples, so that you can see the distribution, so we left it in the design. In the end when we loaded the differentially expressed genes, it was interesting to see the different shapes based on the same values but different distribution, so it did still add value. We ended up adding a save button to save the name of the gene of interest, which then links to open a new page which provides gene ontology information from a reference database. We implemented this for both genes and gene pathways.

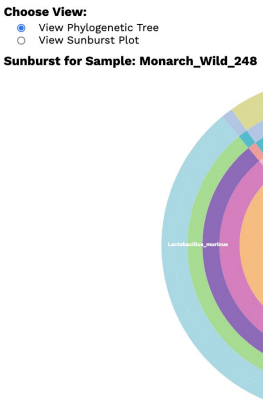
Sunburst Plot

The sunburst plot was not in our original design, but it was something we saw and liked early in the brainstorming process, we just thought that the bars would be better to indicate proportion. However, once the stacked bar chart was working well we decided to try adding it and realized that it added a new dimension to our visualization because now you could compare every level of taxonomy at the same time and focus on one sample. By looking at several samples one after another, you can see how much the samples vary and how they vary. We needed an interface to be able to choose which sample you were looking at and since they were already listed on the screen on the x-axis, we used that as our method for navigation. When you click on the sample name, it brings up the appropriate sunburst. Youjia suggested that we make the text larger on mouseover, and that was very helpful for seeing where you were and making it clear which sample you were clicking on.

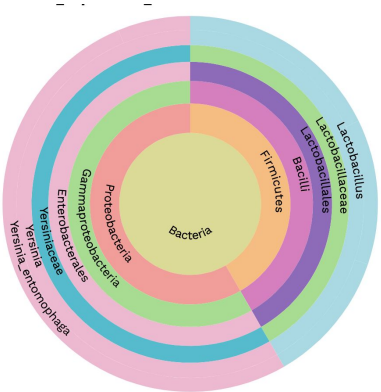
Text on Sunburst



Drawing the text labeling on the sunburst ended up being much more difficult that we anticipated because there is no built in function to place the text between two lines. You have to somehow create a path, or reference from an existing path. We tried straight text first, but it did not look good, got occluded and was difficult to read. We made the text tiny but it just wasn't very readable. The small shapes were too small for text so we took a threshold of 15% and did not show text if the species had less than 15% of the total abundance.



We first tried a method of drawing a curved line that looked very nice, but it created a ton of extra hidden shapes that didn't work well for our dynamic visualization, so we had to try something simpler so that we could still use our zoom on click features and mouseover.



We ended up writing a formula which linked the text directly to the path of the sunburst, moved the text down so that it was inside and placed it on the edge to try to get most of the text vertical. Under the tight time constraints (we didn't start the sunburst feature until Thanksgiving weekend), it was the best we could do. It still has some issues when you zoom in and out, but it is readable and provides value so we left it

in for the final design. In the future, we would like to improve this aspect of the design. Below is a photo of the final version.

Debugging

Debugging ended up taking much more time than we anticipated, and fixing little problems with the interactions between the components took most of the last three days of the timeline, and we are still finding little bugs here and there that need fixed. Here is a list of the known bugs that we were unable to fix because of lack of time at the end.

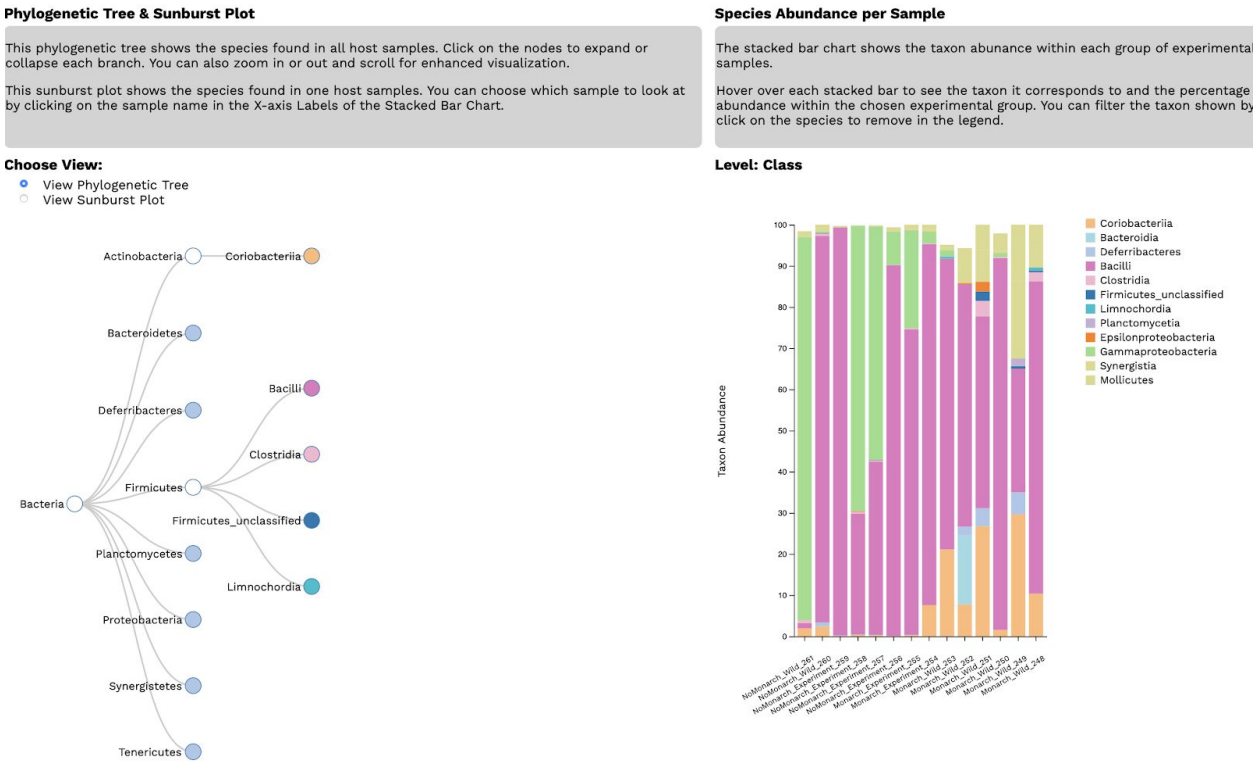
Known Bug List

- Text on the circle stays on the path of each arc, but the text sometimes blocks the ability to click the inner circle to get back out (if you get the curser to look like an arrow instead of a text curser, it will work, but it is not intuitive and sometimes you can get stuck at the bottom of a sunburst after you have zoomed in.
- Better implementation of the text on circle would have all the text that ends up upside down flipped on the bottom so you can read it, and would redraw after the zoom so that arcs that get expanded also get a label, instead of only the original text labels.
- The species level of the tree (when the tree is completely expanded) does not update with color.
- Sometimes mouseover is delayed on sunburst
- Labels in the tree occlude when the names are very long like “unclassified after a family name”

Implementation

We walk through the interaction elements of our data visualization tool.

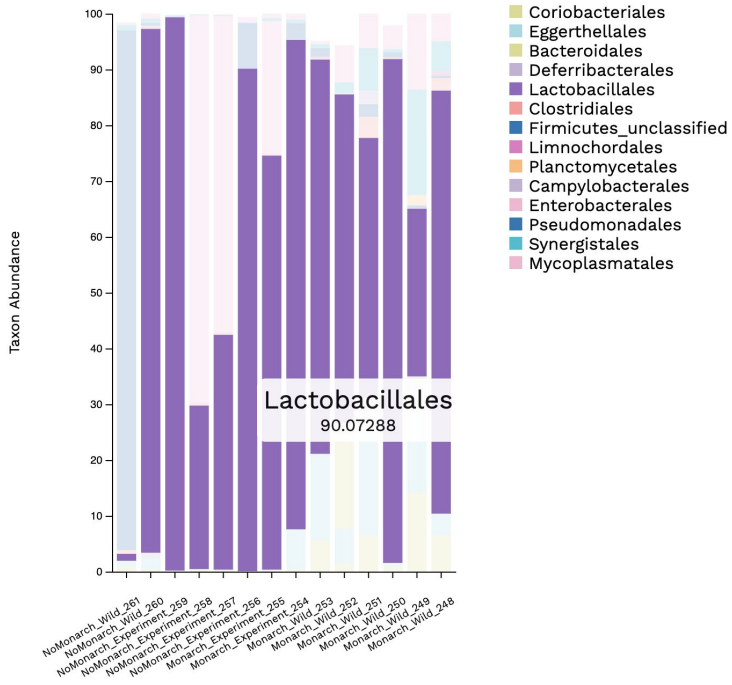
Stacked Bar Chart Updating According to Tree Depth:



The user can click on a node to see further divisions of the taxon. As you increase the depth of the tree by expanding nodes from left to right, the stacked bar chart updates the abundances according to the deepest level of classification displayed on the tree. For example, the deepest level of the tree shown here is Class, and species abundances in each sample are now pooled together into the classes they belong to. This allows the user to examine abundances from different levels of classification, from Kingdom down to Species. Nodes at the deepest level of classification are colored to match the corresponding bars in the Stacked Bar Chart.

Stacked BarChart Highlighting:

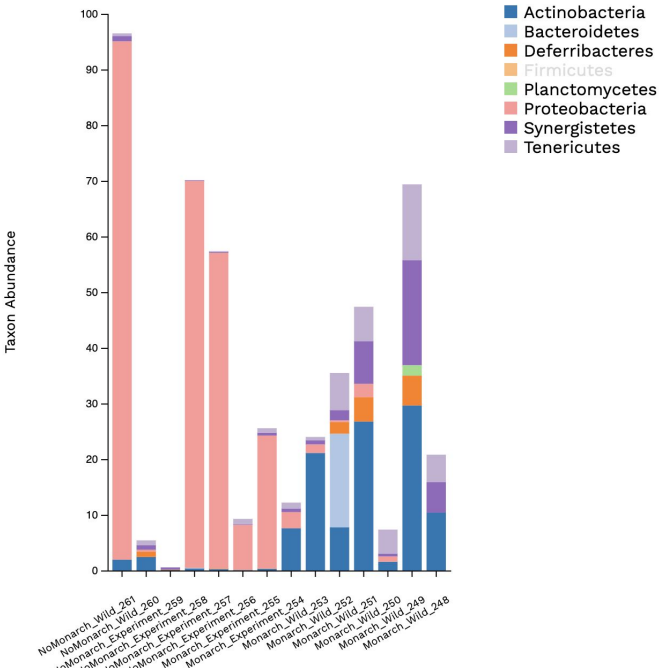
Level: Order



If the user mouses over a column in the stacked bar chart, it will highlight that species across all samples and display an informational tooltip. The tooltip gives the taxon name and its abundance within that sample as a percentage. Since there are far too many colors for the user to pick out by eye, this is an effective way to link the same colors across samples, allowing the user to compare apples to apples.

Stacked Barchart Filtering:

Level: Phylum



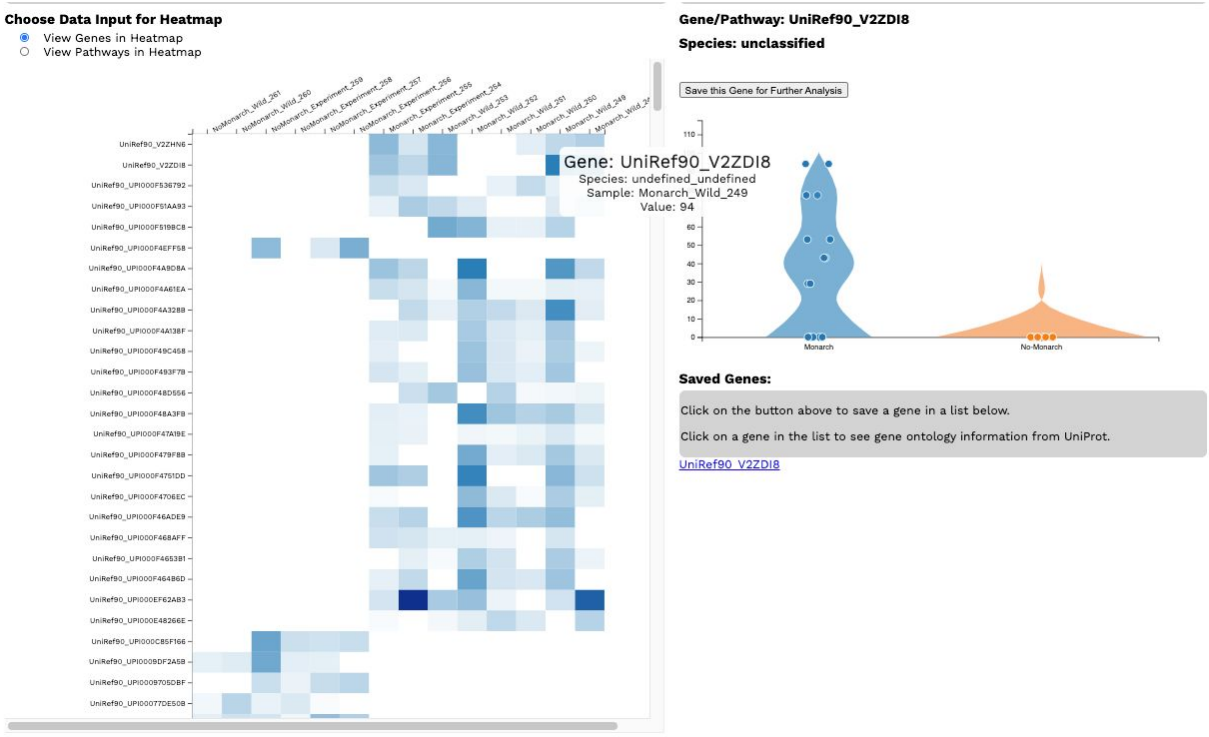
Since the tree may not show all the members of a classification level at once, the user may see all the members of a level from the legend.

The legend is also interactive: clicking on any key in the legend removes the taxon in the stacked barchart.

In this example, we removed Firmicutes abundances from the barchart.

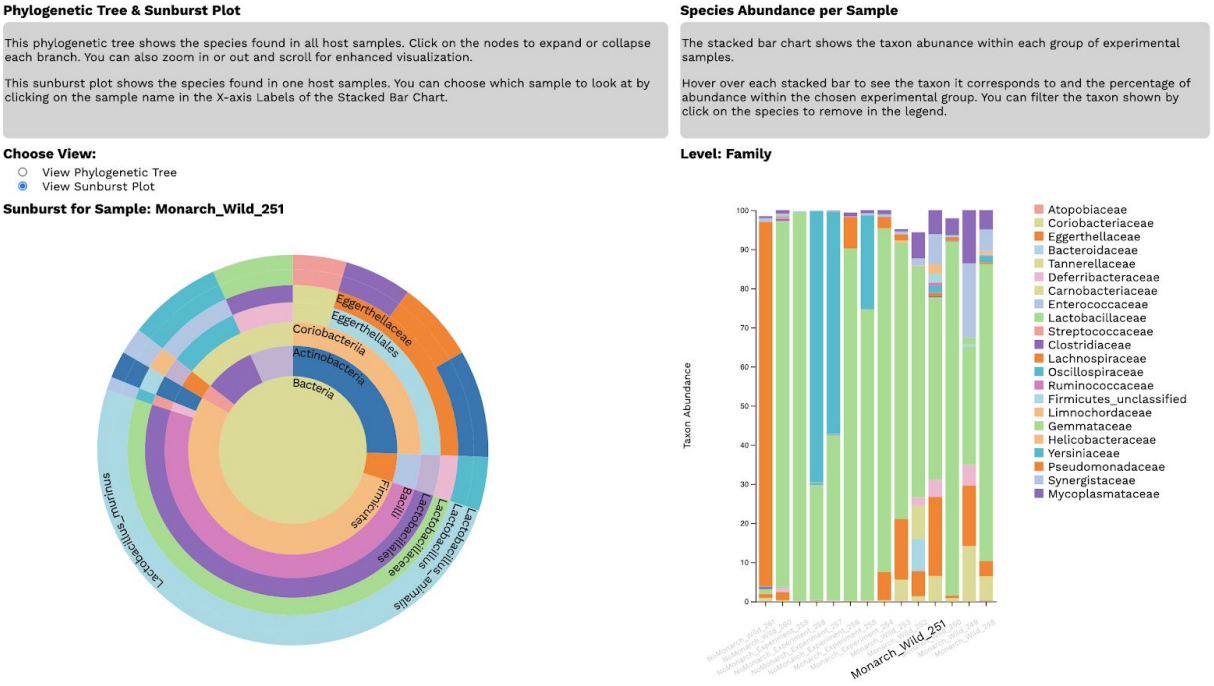
Since we are interested in seeing the differences between the species present in the monarch vs in the non-monarch diet, a researcher might want to remove species that are present in all samples and therefore likely not contributing to the detoxification of the plant toxin, so that the ones that are remarkable are more clearly visible. If the user does not wish to completely eliminate the species from the chart, they can just mouse over the legend and the percent opacity of all the bars corresponding to that taxon will decrease.

<u>Clickable Heatmap that updates Violin Plot</u>:



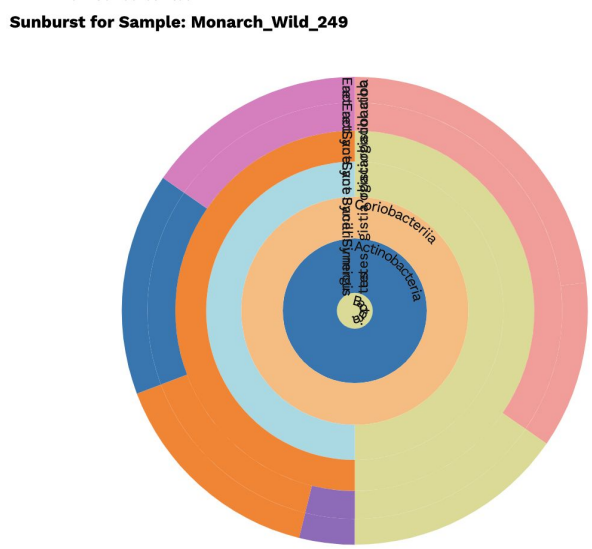
The user can mouse over a cell in the heatmap to see information about a gene along with the sample and its expression level. Clicking on a cell in the heatmap will update the violin plot with a distribution of the expression levels of that clicked gene across samples in Monarch vs. No-Monarch groups. This allows the user to see differential gene expression across the two experimental conditions. If the user wants to learn more about a particular gene, they can click the “Save this Gene for Further Analysis” button, and a hyperlink appears under “Saved Genes”. The hyperlink takes the user to the UniProt website to see gene ontology information.

Sunbursts to Show Abundances in Selected Sample:



The user can look at the abundances of a particular sample closely by highlighting over the x-axis label for the sample of interest. Clicking on the label generates a Sunburst plot of the abundances across classification levels.

Drilling down Abundances by Level on Sunburst:

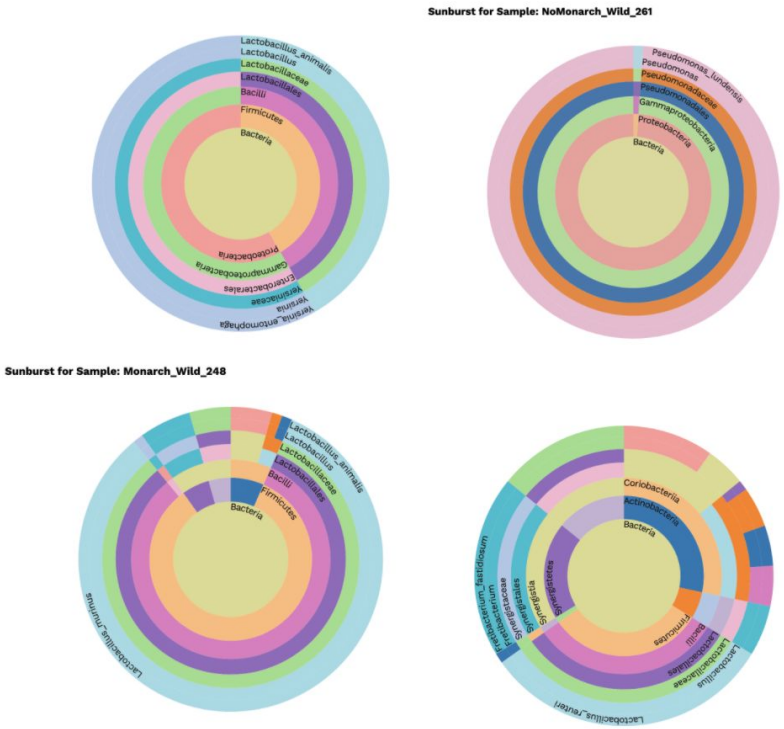


Clicking on any slice shows the abundances of a selected sample across the classification levels.

Evaluation

**What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?*

This specific dataset, the Black-eared Mice Gut Microbiome, is unpublished and a small sample set, so we were not sure how much information we could get from visualizing the data. Even early in the design process, however, we could tell visually that the two different conditions were very different. The Monarch dataset is significantly more diverse at every level of taxonomy. The sunburst plots ended up being a very easy way to spot the differences between individual samples, especially if you compare them side by side.



Our aim was to create a visualization tool that would save the researcher time, so that they would be able to do an initial exploratory analysis of their data quickly, having access to many different views of the data and direct link out to gene ontology and gene pathway information. We originally thought we could have that information on screen, but it is just so dense that we thought the link out was convenient enough and much more complete.

Since we only finished the implementation a few days ago, we only got feedback from two lab members, but they were very excited about it, asked for more features and asked if they could host their own data live so that they could provide links to outside people, which I thought was a very nice complement to the usability of the tool.

Technically, there are many things we could do to expand and provide more functionality, as well as fix some of the integration bugs. First, several people asked for even more ways to subset the data in the stacked bar chart, by either clicking on a single node in the tree, or clicking on Firmiculites and having everything else go away and keep the filtering active as you drill down the tree. We could also rescale the axis when you use that feature so that you can see low abundance species more clearly. The heatmap also could contain more data, and could filter based on species in the tree.

A difficult but valuable addition would be if we could create a phylogenetic tree that could be created not only from the reference database, as it is now, but could map the unclassified species into the tree based on their DNA sequences. That would be extremely valuable in our lab, as we look at

70%-80% unclassified species, because we are looking at wild animals, and their microbiome is not well characterized.

Overall, we feel we created a useful tool and achieved many of our original goals. It was a short timeline, but as a prototype, it is very useful.