Tuberculosis (TB) Dashboard Process Book

By: Monét Norales and Sudheshna Bodapati

Overview Summary: Our initial project proposal aimed at creating a location where researchers in biology or bioinformatics could go to find localized information on tuberculosis. Because the information and data used are from various government databases and research papers, we hoped that the created dashboard would be able to show the top mutations of tuberculosis. For the visualization, we wanted to be able to display and easily identify the mutation locations on the gene, the type of drug resistance they are associated with, and what mutations were most common based on the regions where the samples were taken from.

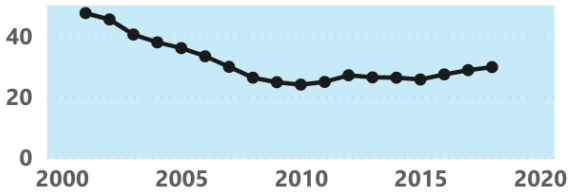
Motivation: The motivation and inspiration for this topic came from our interest in the field of bioinformatics. It is also partially motivated by a personal interest due to genetic or racial connections as about 65.5% of those who have latent TB are of Asian and Hispanic/Latino descent. There are some personal experiences along with support from each other during this project and pursuing the topic from the beginning of initial idea proposals. Also, because there are still a large number of deaths that are still caused by tuberculosis and the drug resistance is becoming something that is being increasingly researched, we wanted to make something useful.

Intended users: As the community searches for a refined drug resistance mutational database to understand its impact on both genomic level and population level they should be able to find this dashboard/website to be a decent reference for the purposes of finding drug-resistances links to specific genes. This could also be used for a more simple population health analysis and overview. Since mutations are always a possibility, there is an ever present potential for increased difficulty in noticing and addressing trends in mutation rate or comparisons between the various mutations, related gene locations, and their influences. This website is to encourage the research community to look more into TB and gain a better understanding of the main mutations that occur and where each one is most documented based on location.With this, it could decrease the interest in further research if the data is difficult to manipulate and investigate.

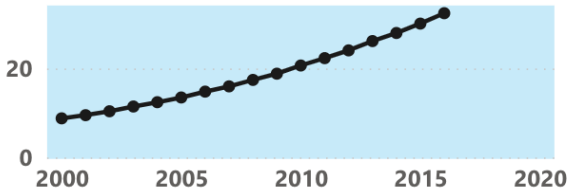
Related Work: We had found a few dashboards as a reference. We also found that of the dashboards on WHO and CDC websites however, we noticed that none of them had very many that addressed the relationships between genes and mutations. There were a couple that did focus on a specific type of drug resistance, but not about the top ones. The dashboards present on WHO's website are shown below. We got some idea of what types of visualizations from some of these dashboards:

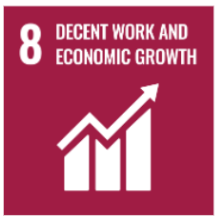
Indicators in the Sustainable Development Goals Associated:

Prevalence of undernourishment
(% of population)

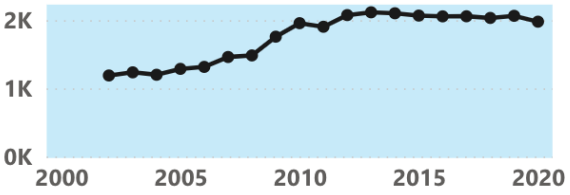


Access to clean fuels and technologies for cooking
(% of population)

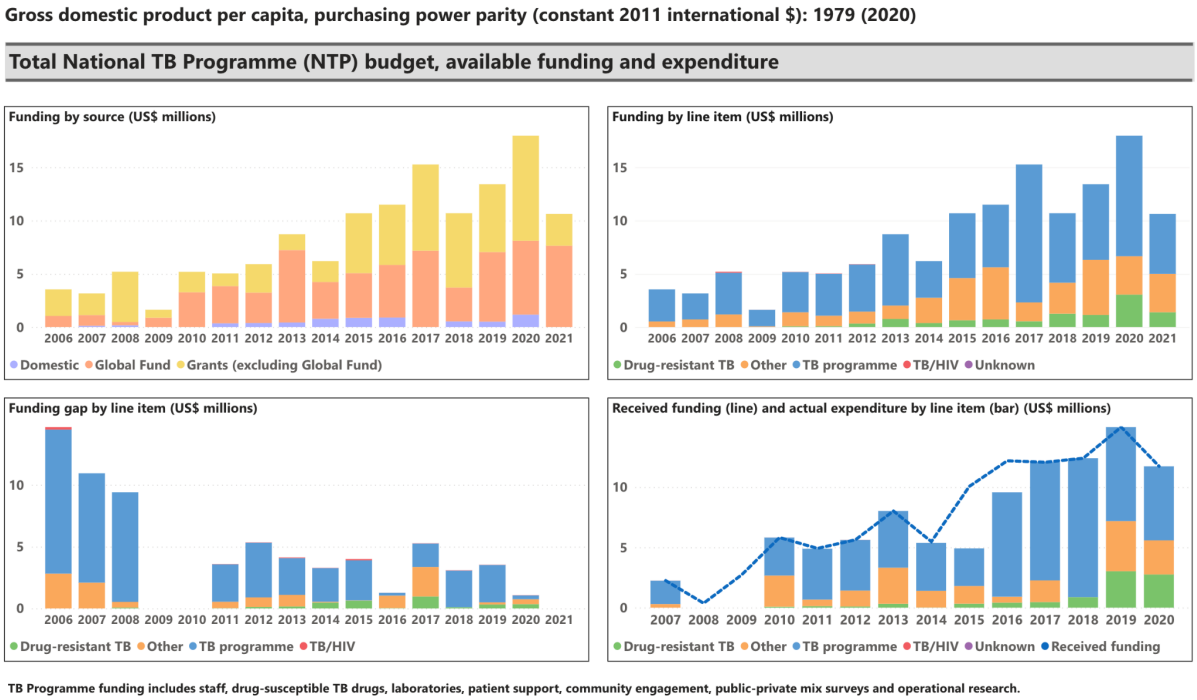




GDP per capita, PPP**
(constant 2011 international $)



Financing got TB prevention, diagnosis and treatment:



TB country, regional and global profiles:

Tuberculosis profile: United States of America

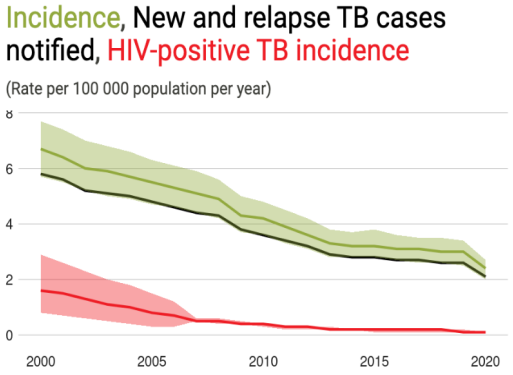
Population 2020: 331 million

Estimates of TB burden*, 2020

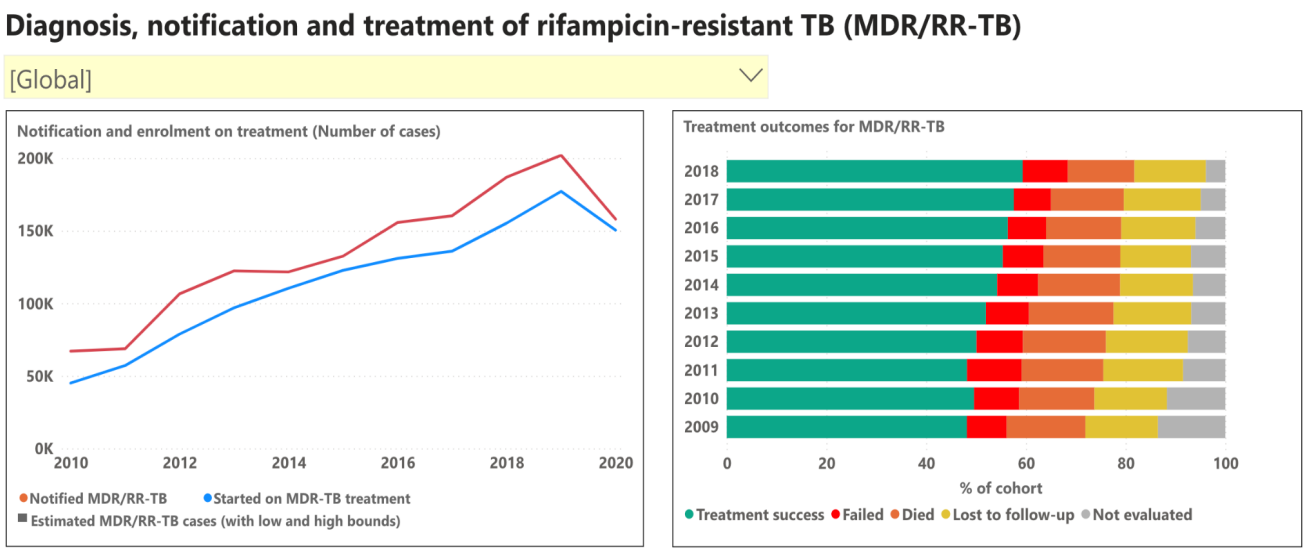
	Number	(Rate per 100 000 population)
Total TB incidence	7 900 (6 700-9 100)	2.4 (2-2.7)
HIV-positive TB incidence	380 (310-450)	0.11 (0.09-0.14)
HIV-negative TB mortality	550 (550-550)	0.17 (0.17-0.17)
HIV-positive TB mortality	68 (46-93)	0.02 (0.01-0.03)

Universal health coverage and social protection*

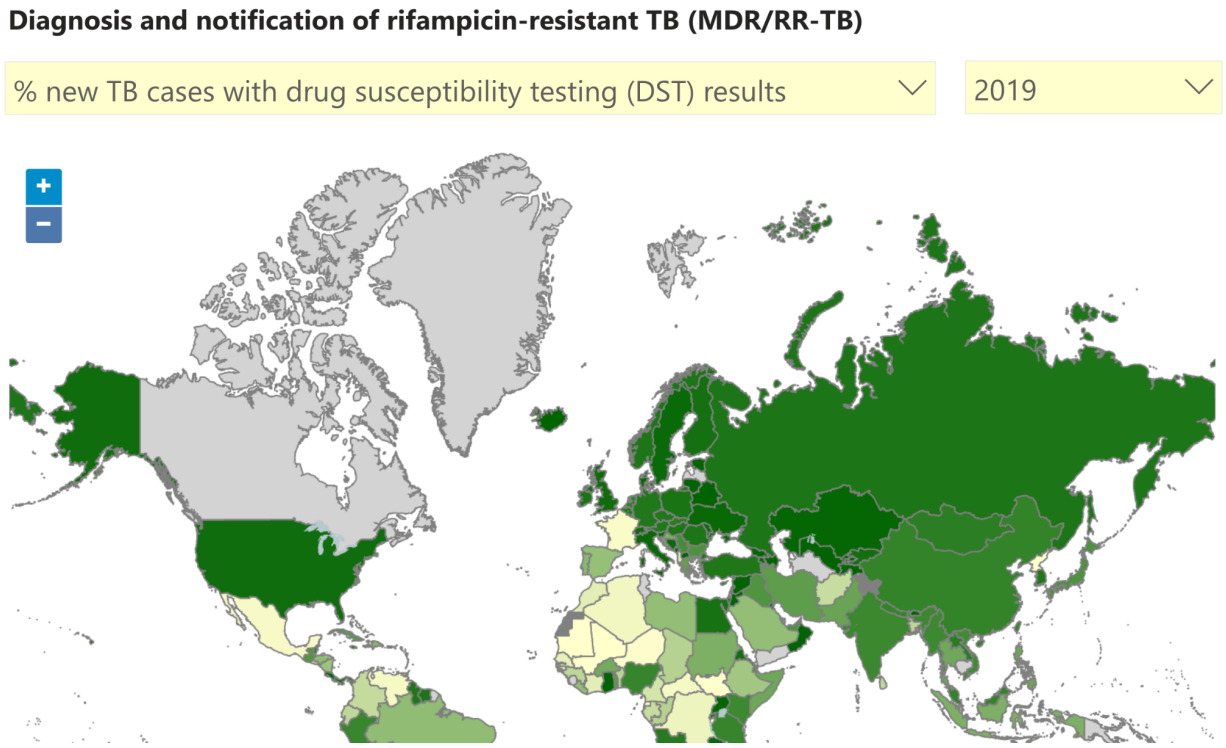
TB treatment coverage (notified/estimated incidence), 2020	87% (75-100)
TB patients facing catastrophic total costs	
TB case fatality ratio (estimated mortality/estimated incidence), 2020	8% (7-9)



Diagnosis, notification and treatment of rifampicin-resistant TB:



Maps on the diagnosis and notification of rifampicin-resistant TB:



We also had wanted to include some of the method of visuzliaztion that were talked about in class and decided to include a force graph.

Questions: Using this dashboard, we wanted to be able to show the top mutations of Tuberculosis drug-resistance and the genes and gene positions associated with it. As we progressed thorugh the project, we kept getting more and more ideas on what data to show or how to show it. Due to the time constraint and that D3 is a new language for both partners, the top ones were implemented and the rest were saved for future implementation ideas. In the course of our analysis, we considered the aspects of what was given as information in the data set as well as what would be directly showing ewhat we haunted to investigate. We also considered what would be extra information and because the data bases come from different locations, how including such information might make the resulting visualizations less reliable or accurate due to the different origins and no real feature identifiers to merge them on.

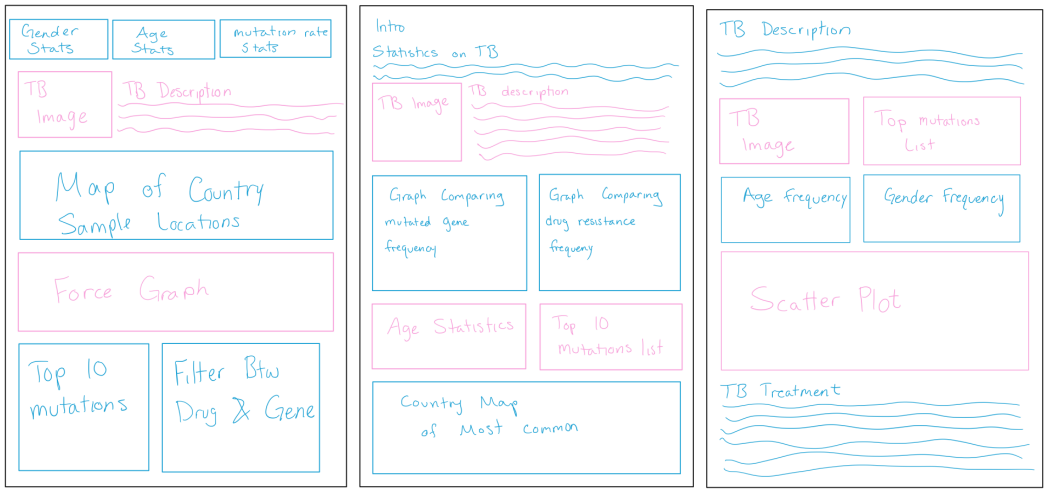
Data Source: Our first set of data was downloaded from WHO's tuberculosis website. It contained various features including: country, year, age-group, sex, risk-factor, etc. This data was initailly included in the page designing. It was later changed as we changed data sets. The dataset we used for this was found on the DRAGdb website. Because our goal was around creating a dashboard for tuberculosis mutations, this dataset contained more of the information that we needed. It had features — gene name, drug name, nucleotide position, and nucleotide mutation

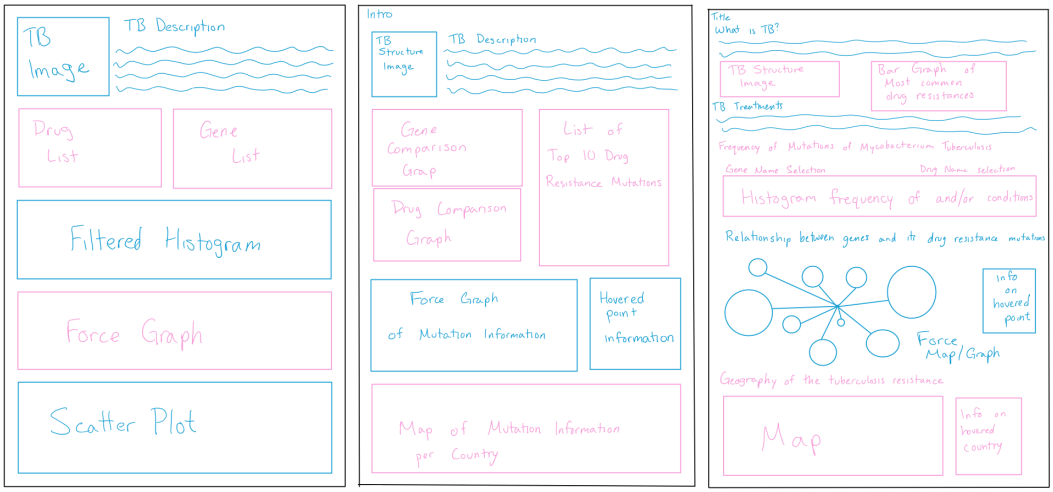
— which were more in line with what we wanted to be able to visualize and what information we wanted to convey to the users.

Libraries: As for libraries, we used the javascript library with Geo.path for the world map. When creating the node/force graph, we used a force directed algorithm and had the forces between the elements set to be attracted to the center of gravity and repel one another.

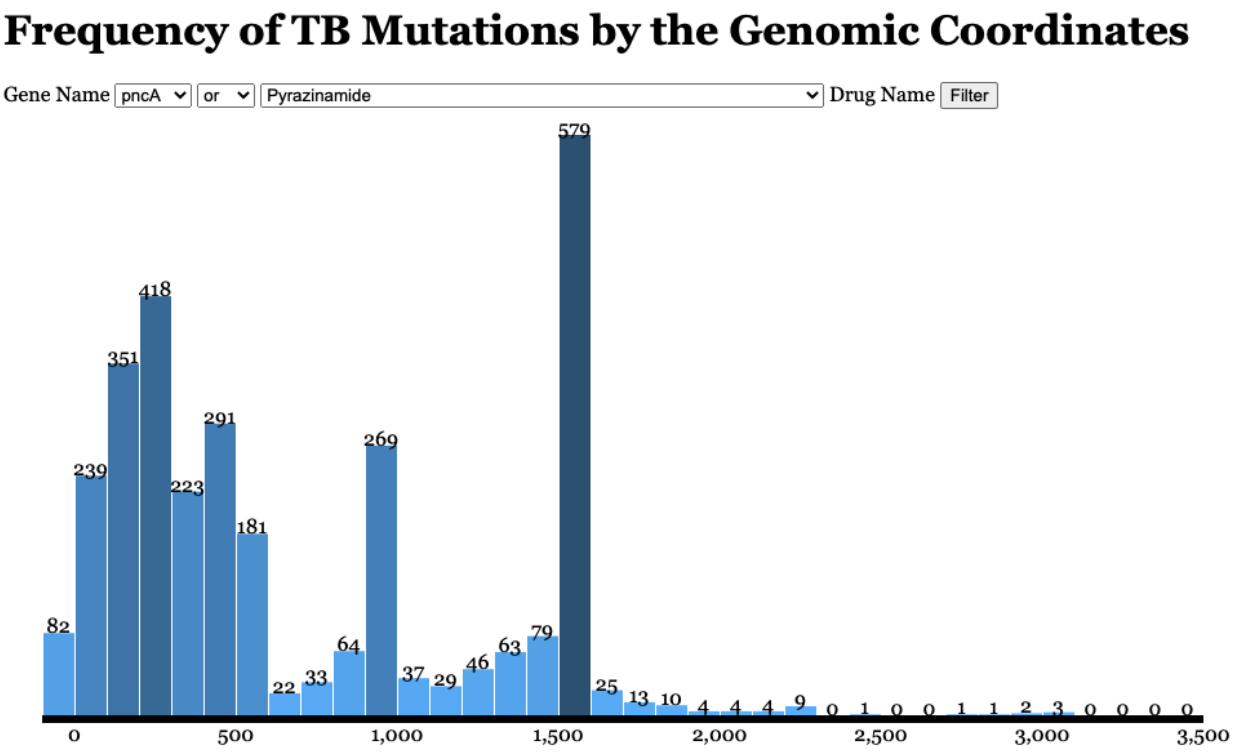
Exploratory Data Analysis: We had initially looked at simple bar graphs, pie charts, histograms, and scatterplots for our visualizations. However, as we looked more into the avaliable data, we thought of more dynamic and intersting ways to show the same data in better ways. Exploring more into the data aslo helped us determine what would be worth focusing on visualizing and what would be the most effective ways to display such information. While doing this, we were testing our abilities and skills at coding and learning new coding methods in a short period of time. This lead to the investigation of visualizations such as the implemented world map and the force node graph. From these we were able to more readily find the answers to our initial questions while also helping to stimulate ideas for continuing and what could be done later on.

Design Evolution: Below are some of the webpage designs that we originally came up with along with the final layout. We did not deviate too much from the original proposal in terms of what we wanted to show and the general design set up, however, we did have to adjust for the change in data set later on. The below design drawings are in order from left to right as we adjusted and reorganized the webpage set up.

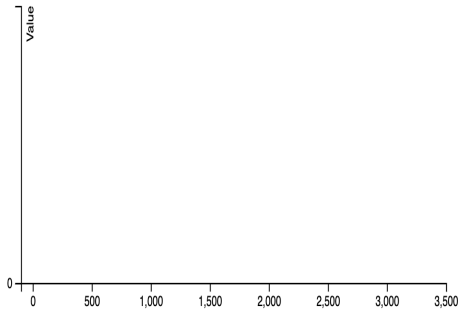


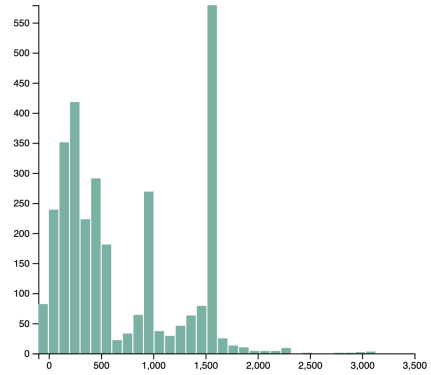


Implementation: Our interactive visualization include a bar graph where the user can select the gene name and/or the drug name and the graph will update to show the frequency at which the mutation of resistance occurs and at which location on the genome. During the implementation process, we encountered many difficulties and complications for this chart. When creating the graph, there were struggles with adding a filter as well as difficulty rendering the bars after adding the filters.

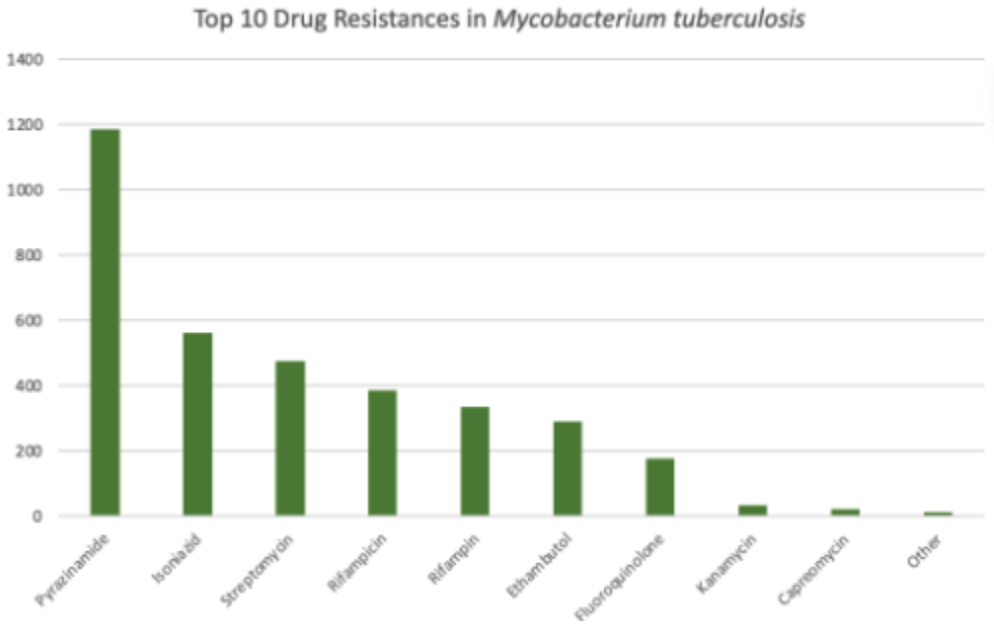


Before doing this, we had tried to create a bar graph to show the most common drug resistances found in tuberculosis based on the samples taken form the data set. We had more difficulties in doing this, as seen below.





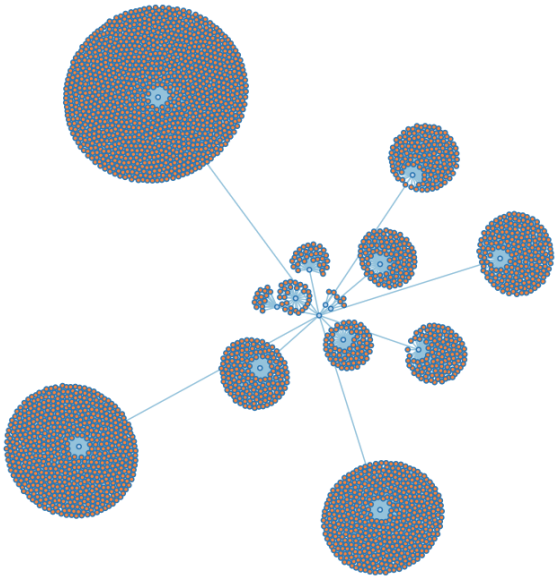
Because we were unable to show the data in a way that we had intended, and ended up using Excel to create the static graph at the top of the web page to display the intended information. The result is the graph below.



The next of the interactive visualizations was the force node graph. The grap itself is collabsible and each individual root node is representative of a gene while the leaf nodes are the drug resistances with the positions that ther drug resistance is correlated or connected to. Due to this relation, the size of the node clusters are indicative of the number of mutations associated with a specific gene. This one was particularly difficult because of the need to assign a root node and then the drug names to it. We also needed to make adjustments to the sizing on the node graphs. Because of the formatting and because the d3 language is not something we are particularly

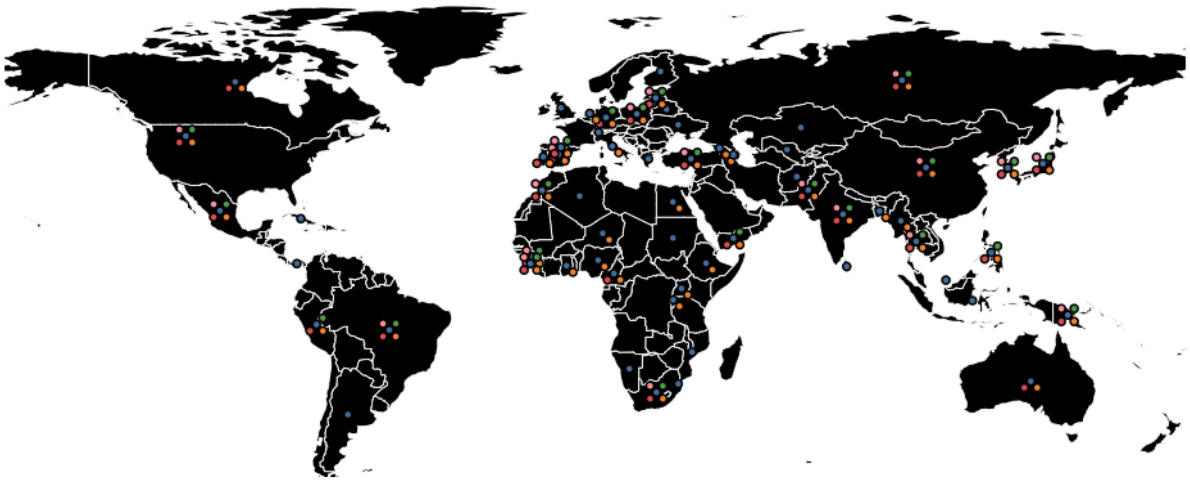
confident in, we had trouble when appending the svg element to the nodes in thew graph. While doing that we were able to get assistance from others with more experience in the d3 coding language and were successfully able to render the visualization.

Relationship between Gene and Drug Resistant Mutations

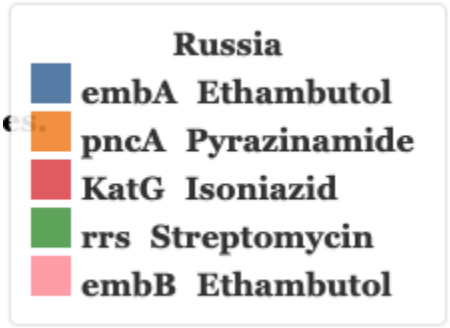


The next and final visualization that we were able to complete was the creation of a world map. This is one of the more complicated visualization types that were discussed in class that we wanted to try. When creating the world map, we found that appending and displaying the data points for the top five types of drug resistance required more involvement that we had initially thought. Below what the initial map of the world that we were able to render.

The final resulting graph that includes the points for the top drug resistance types for each avaliable location. And an example chart of what will show up if one of those locations is hovered over with the mouse.



An important note to be aware of is that the colors of the points are not consistent between each area in determining what drug resistance is there. For example the gene drug combination embA–Ethambutol is blue for Russia as seen in the below screenshot, however, another combination may be shown as blue for another location on the map as seen for Pakistan below.



Evaluation: From using our visualizations we were able to find that the most common drug for resistance to be found for in this dataset was Pyrazinamide and the most common gene to have mutations in it was pncA. We were also able to learn which areas are more likely to have a specific drug resistance. Also becasu we are able to get these answers from the visualizations, it is safe to say that the final version of our visualization works as intended. We have thne nodes moving and all of the interactive features function as intended despite the complications and issues that arose while implementing them.

Conclusion: When completling this project, we found that it was much more difficult than we had anticipated, but not so much so that it was unbearable. We enjoyed being able to apply the different visualization techniques learned in class as well as learning about and getting to gain

experience in a new coding language. We also enjoyed being able to see the application of visualization and how itcan be used to make data more interpretable and accessible for a variety of people depending on the intended audience.

Future: Though we considered this a successful implementation, there is always room for improvement. To improve this dashboard, we could add more about specifics, maybe including other visualizations to show the correlations or connections that might appear between things like age and gender. Also instead of just adding more visualizations, we could do more to better format the webpage to change the asthetics more by creating clearer barriers between sections or recoloring the world map. We could also add a zooming fuction on the world map to beable to more clearly see each region that the samples originated from. In the consideration of expanding the data used, we could include more on specific genes and trhe specific nucleotide regions in which the mutations take place, create a filtering function to be able to see the top drug resustance mutations per region for each year the data is collected for it, or using this information and maybe more data collected from other papers and databases to identify the conserved regions within different genes to find a target for vaccine development or effective treatment.

Peer Evaluation:

- Review of Sudeshna: An excellent project partner. She was very flexible when creating meeting times and completed all of the work she committed to within the agreed times. Showed amazing effort when trying to understand the new language and often set aside other projects in order to prioritize this one.
- Review of Monét: Good team member and very cooperative. Met often worked together in groups. We were both struggling to develop code since we are both not familiar with the javascript but we worked together for long hours and made it work. Very understanding and cooperative.