Process Book Submission for Data Visualization CS - 5630/ CS - 6630 Final Project

Investigating Overlap and Segregation of Variants between ClinVar and LOVD Databases

Project Proposal

Data

The data for this will come from ClinVar and from six LOVD databases. The LOVD databases that will be included are Global Variome shared LOVD, Human Variome Project, Mitochondrial Disease Locus Specific Database (MSeqDR-LSDB), Brazilian Initiative on Precision Medicine (BIPmed) SNP array database, the BIPmed whole exome sequencing database and the Cincinnati Children's Hospital Medical Center (CCHMC) database. Genes from three disease categories will be used, Metachromatic Leukodystrophy (MLD), Metabolic Diseases (includes several diseases, but all are metabolic in nature) and Severe Combined Immunodeficiency (SCID). MLD and the metabolic disease genes are the only ones known to be associated with the given diseases, and associations are very clear. SCID is less well understood genetically, so we chose to use the New York NBS panel list of 39 genes. The genes per disease are as follows:

- Metachromatic Leukodystrophy (MLD):
    - ARSA
    - PSAP
    - SUMF1
- Metabolic Diseases:
    - ACADM
    - ACADS
    - ACADVL

    - CPT1A
    - CPT2
    - ETFA
    - ETFB
    - ETFDH
    - HADH
    - HADHA
    - HADHB
    - SLC22A5
    - SLC25A20
- Severe Combined Immunodeficiency (SCID):
    - CORO1A
    - CD247
    - ATM
    - LIG4
    - DOCK2
    - PRKDC
    - DCLRE1C
    - CHD7
    - NHEJ1
    - GATA2
    - TBX1
    - IL2RG
    - MTHFD1
    - RAC2
    - IGHM
    - ADA
    - IL7R
    - MTR
    - CD3G
    - BTK
    - AK2
    - JAK3
    - RMRP

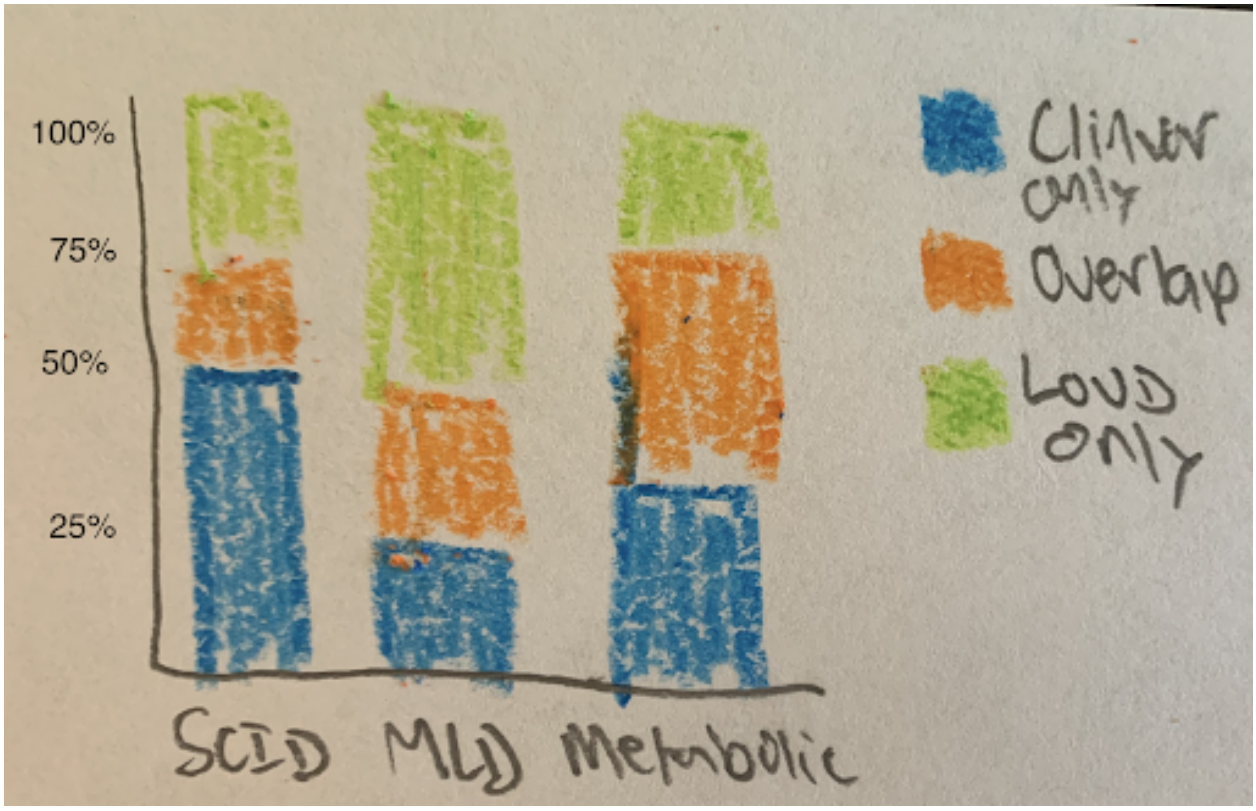
- STAT5B
- CD40LG
- CD3D
- RAG1
- SLC46A1
- ZAP70
- WAS
- CD3E
- RAG2
- DOCK8
- PNP
- DKC1
- PTPRC
- FOXN1
- NBN
- BLNK

Data Processing

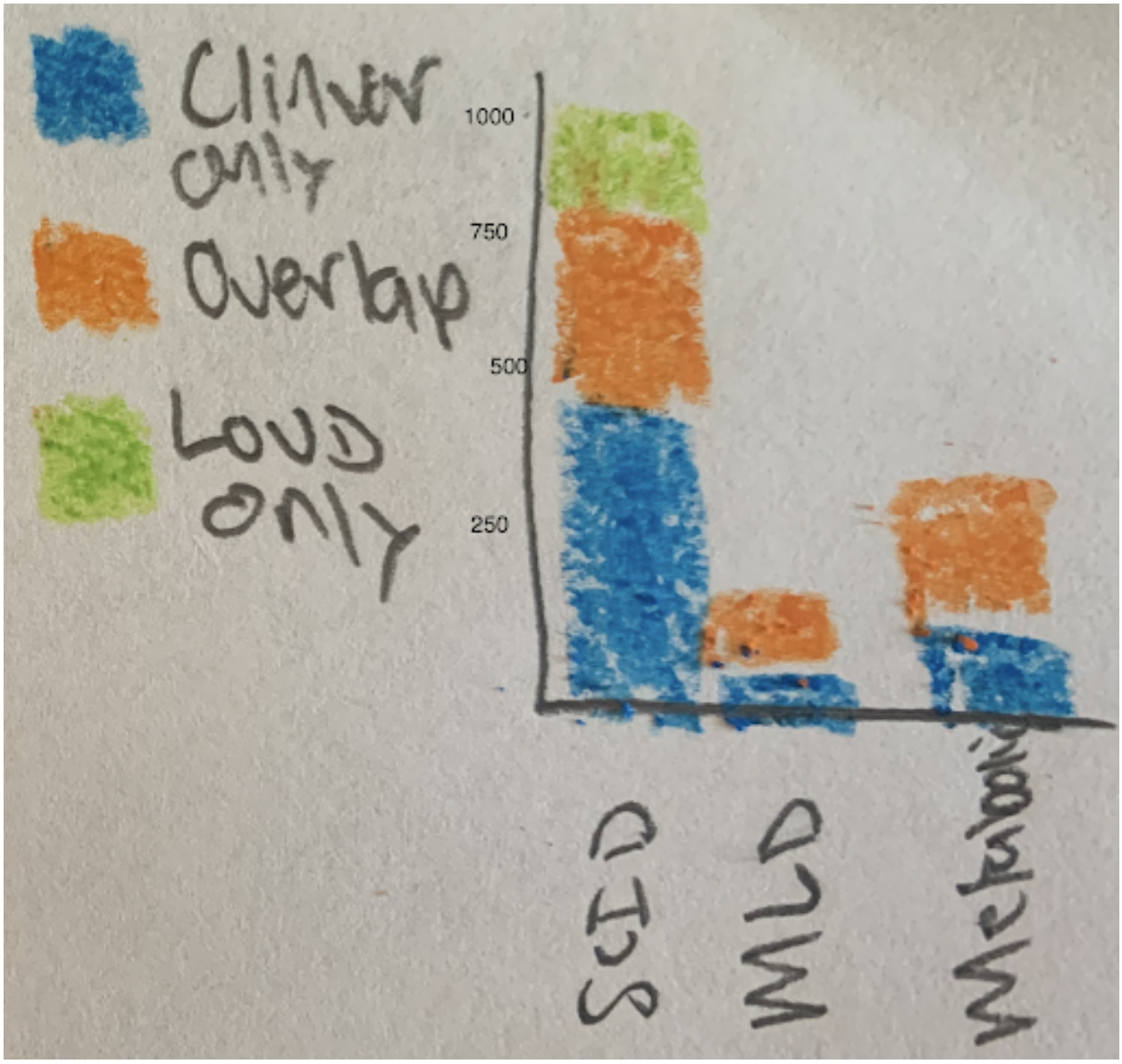
Obtaining and cleaning the data will take significant work. The data will be processed using the python scripts in the project directory in the Eilbeck lab GitHub page. The variant information from ClinVar will be parsed from the XML file that can be found on their FTP directory: ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/ The variant information from the LOVD version 3 databases (all except CCHMC) will be obtained through web scraping pages with tables containing the variant information. The CCHMC website is still in LOVD version 2 format and must be scraped one page per variant entry. All information is saved in csv format. The scripts will obtain the variant information and normalize all variants to HGVS format. Variants that cannot be normalized automatically will be stored in separate files. We plan to write a bash script to parse the gnomAD variant call format (vcf) file to obtain frequency information. We also plan to write a python script to put the information into a SQLite database so that it can take less space and be parsed more easily.

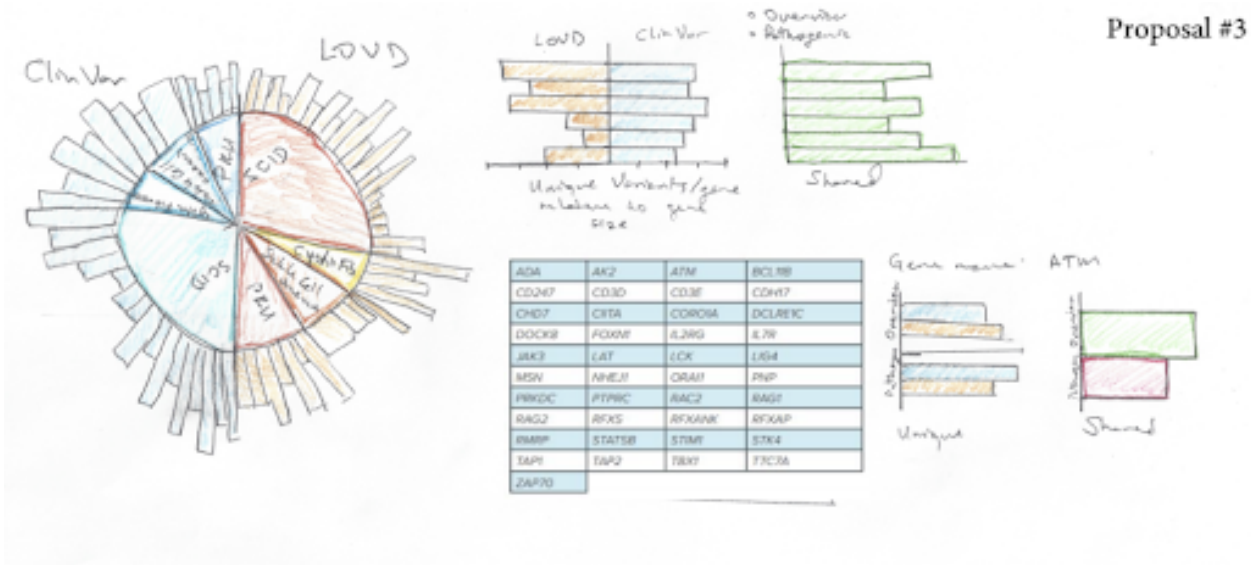
Visualization Design

On page load the page will show a stacked bar chart showing the percentage of variants per disease that are in both ClinVar and LOVD databases, ClinVar only or LOVD only.

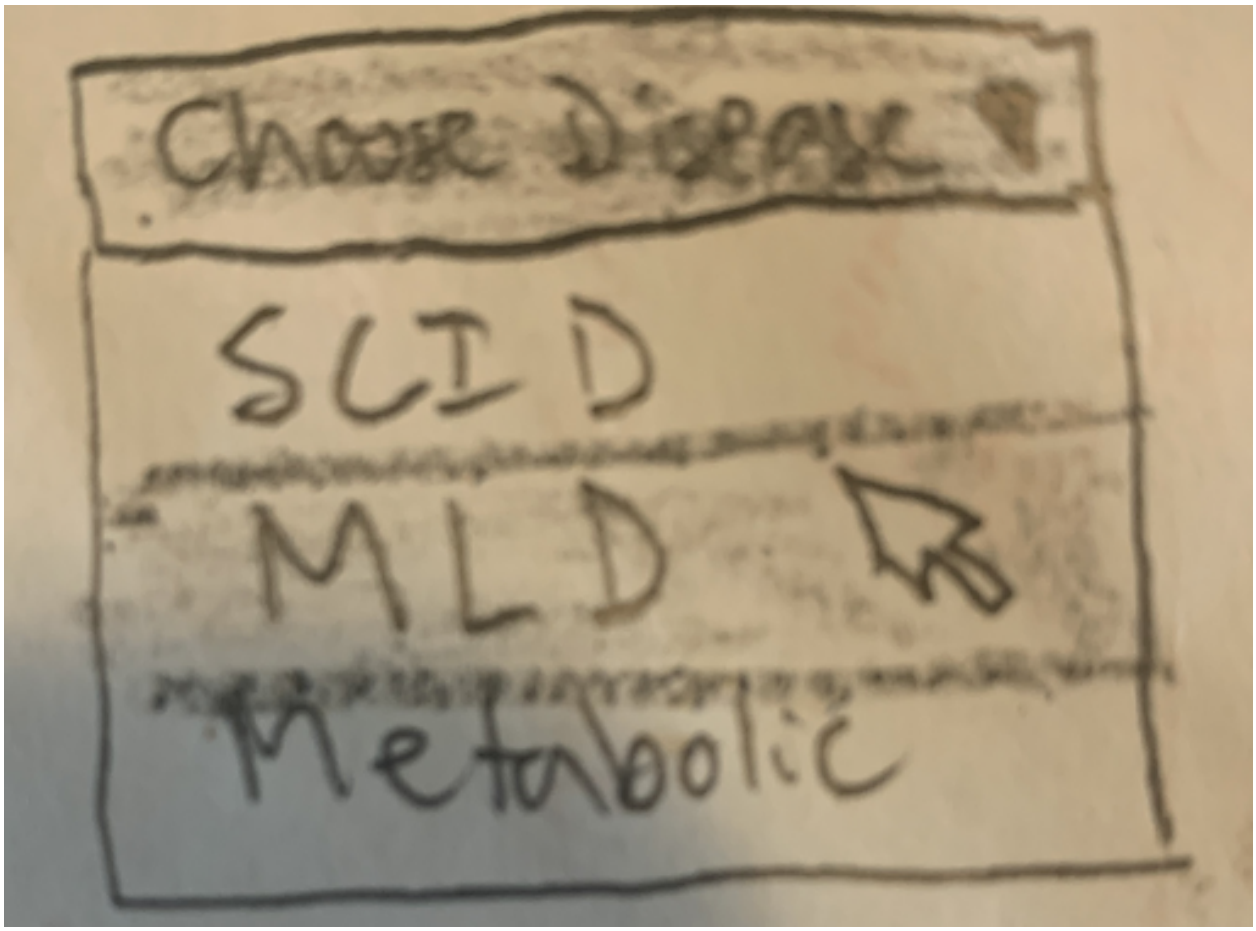


There will be a button to tell the figure to switch from percentage to total variants.

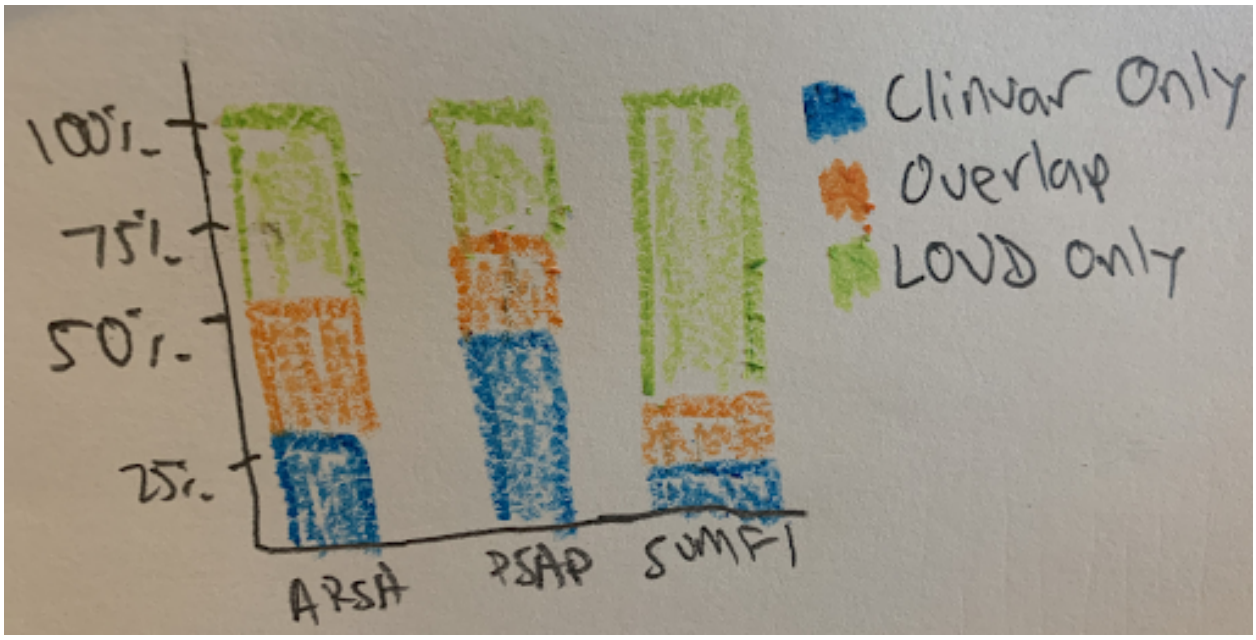

Clinver only
Overlap
Loud only
1000
750
500
250
SCID
WILD
Wrabbit



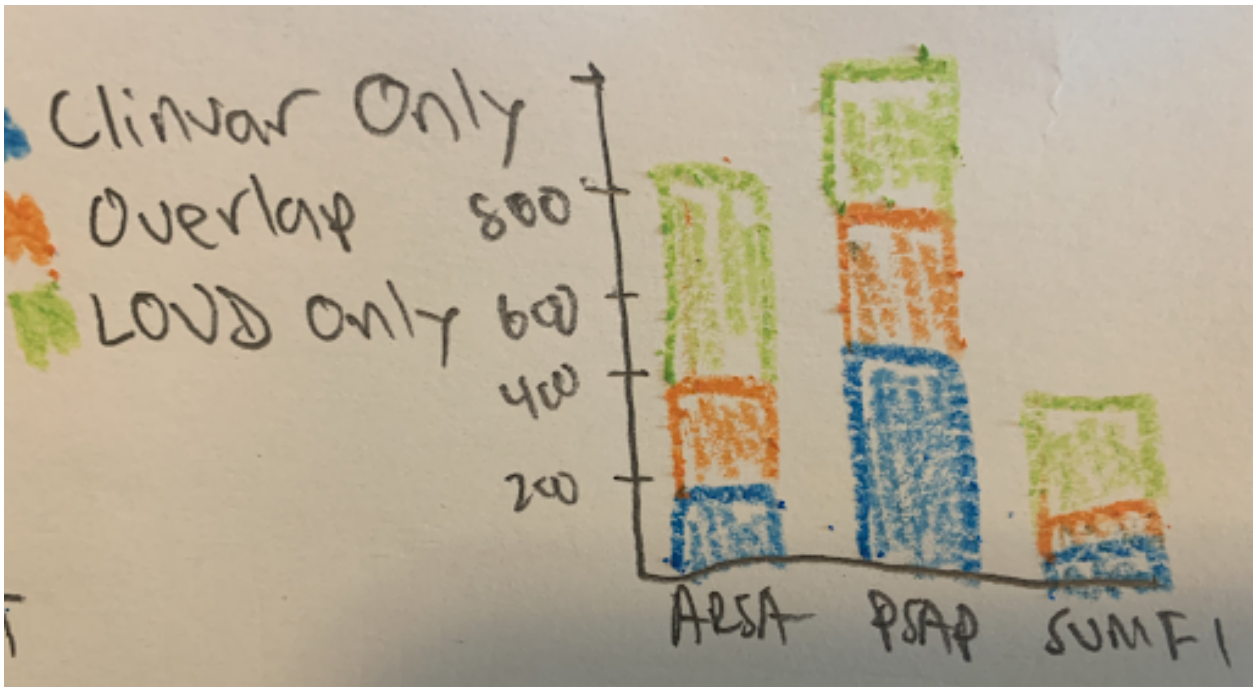
On page load we would also like a dropdown menu to be present to allow the user to select which disease's information will be shown. The options are currently MLD, Metabolic Diseases and SCID.



After choosing a disease, stacked bar charts will show the overlap between ClinVar and LOVD databases per gene.



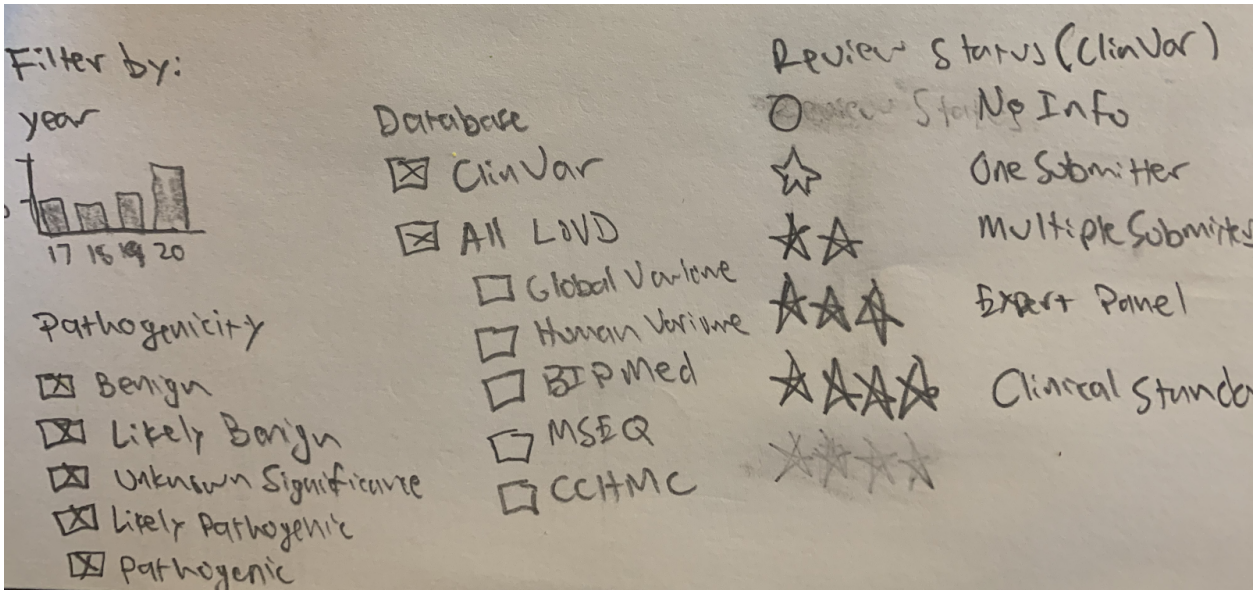
Similar to the stacked bar chart for disease, this one will have a button to click and switch to total variants instead of percentage of variants.



In addition to these graphs, another set of stacked bar charts could be used to show what type of variants these are.

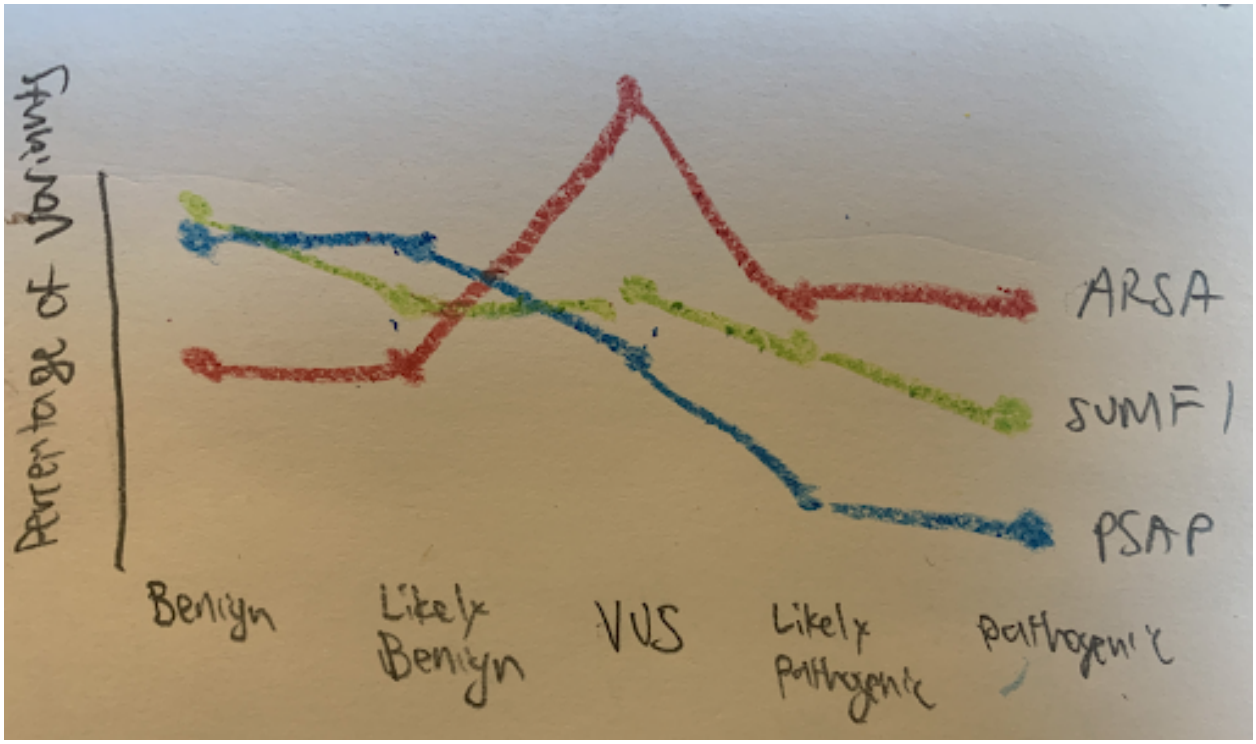
This, like the other bar charts, could be changed to percentage rather than total counts.

Along with drawing these, there would also be options for filtering the data. By default, it will show information about all of the variants. I would like the viewer to be able to select the filter based on several criteria. One of them could be the year that the information was last changed. ClinVar has only been around since 2012, so no variants are older than that, but I thought it could be changed so that you can show only variants from the last two years. Another filter criteria could be pathogenicity. ClinVar and LOVD use the same pathogenicity scales (with slightly different wording). The degrees of pathogenicity are benign, likely benign, variant of unknown significance, likely pathogenic, and pathogenic. The users could also filter the ClinVar variants based on the review status (star level) from clinvar. These are as follows: 0 stars - No submission information provided, 1 star - Single submitter or multiple submitters with conflicting interpretations, 2 stars - multiple submitters with no conflicts, 3 stars - determined by expert panel, 4 stars - clinical standard for diagnosis. Finally, it would be nice if the viewer could filter based on the database. They could select only variants present in both ClinVar and LOVD, variants in only ClinVar, variants only in LOVD, or variants only from a specific LOVD database. Ignore the poorly erased line in the figure below (crappy eraser).



Another graph that could appear on the page after selecting the disease is a line chart showing the proportion of variants for a given gene that are of each pathogenicity.

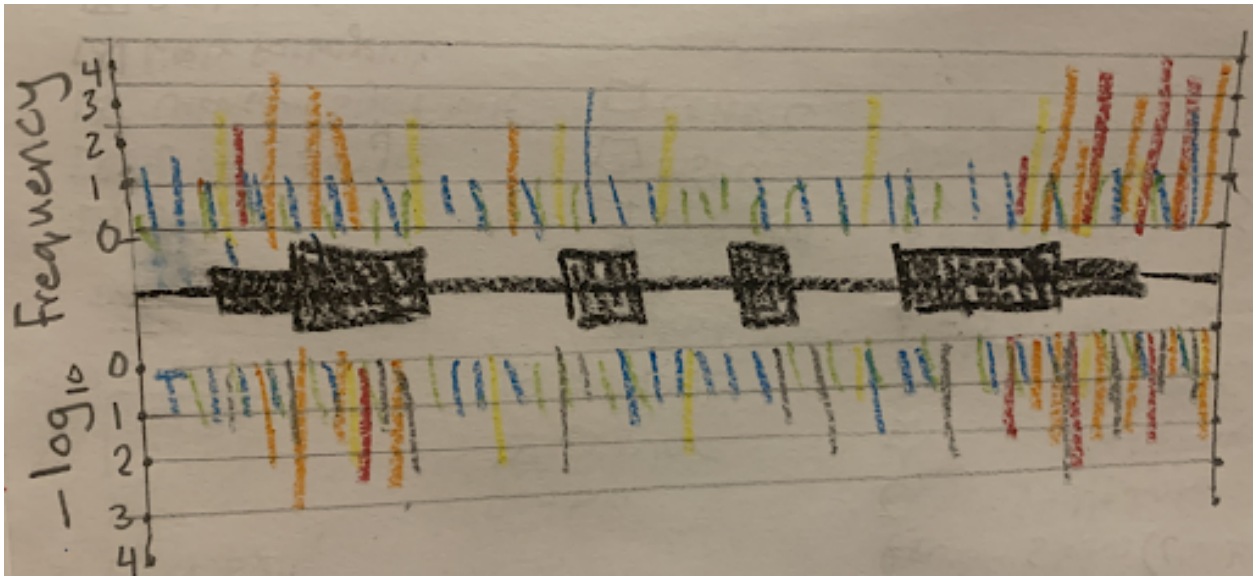
We would like to make a second page that shows visualizations for one gene at a time.



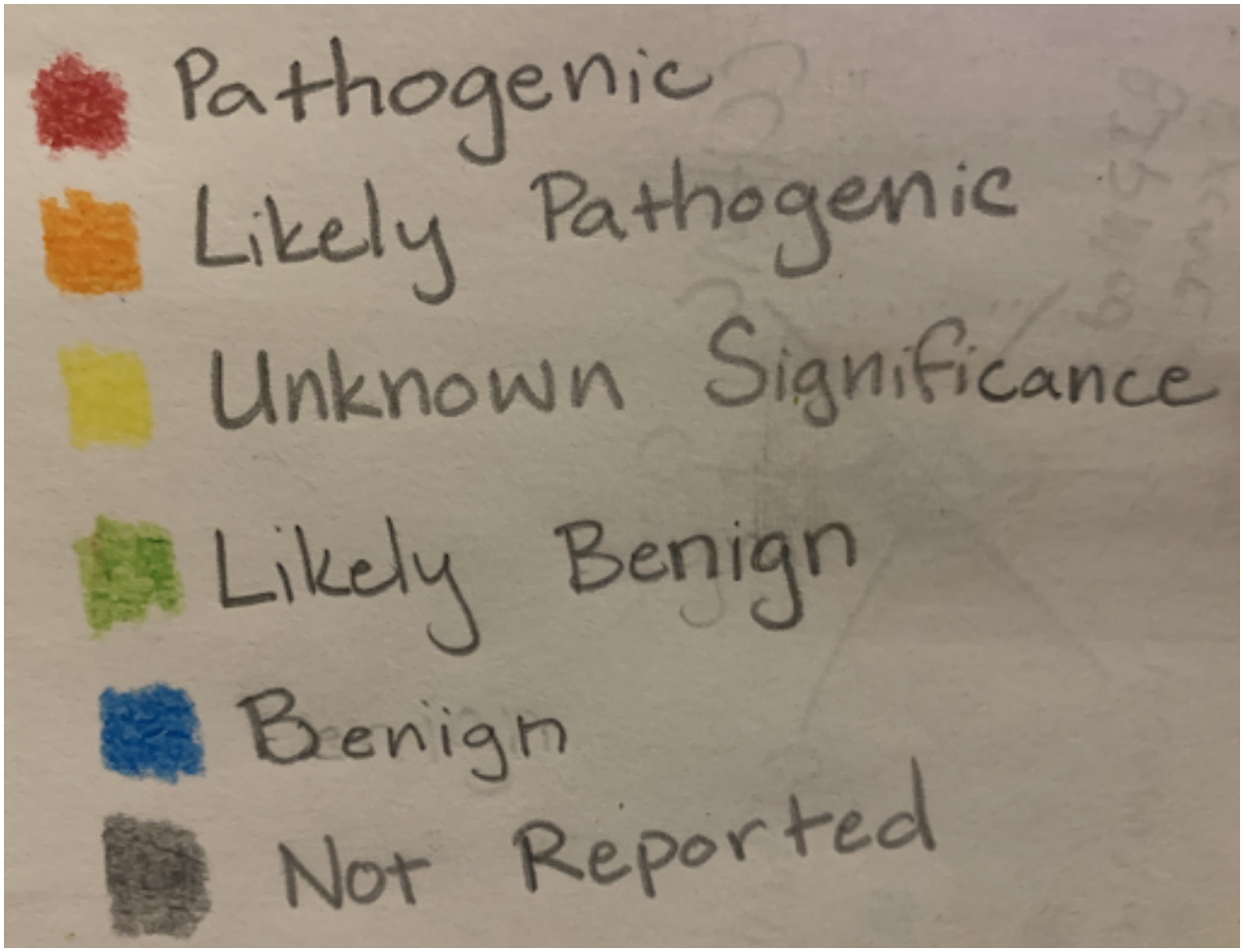
Due to being drawn in crayon, the picture shows labels next to the lines. Ideally, the name would appear on hovering over a line. MLD has only three genes associated, so three separate colors can be used. However, with a disease like SCID, the display of 39 genes would simply be too much. Instead, the color could be based on the number of variants for the gene with a scale on the side. All lines would be at 50% opacity (maybe less) to allow for viewing all lines at the same time.

A second page could be made with information pertaining to a specific gene. Alternatively, this info could be shown at the bottom of the first page. A link from the original page can take the viewer to the second page. On the second page a drop down menu can be used to pick a single gene of interest.

On the second page a chart can show variants' location within a gene. The representation of the gene will follow that of the University of California Santa Cruz (UCSC) Genome Browser. The thin lines represent intronic regions and intergenic regions. The thickest regions represent exons. The regions of medium thickness represent untranslated regions of exons. The variants from ClinVar will be represented on the top of the gene representation and the variants from LOVD databases will be represented on the bottom of the gene representation. Each variant is represented by a line. The height of the line represents the frequency of the variant in the global population on a negative log base 10 scale. The color of the line represents the reported pathogenicity.

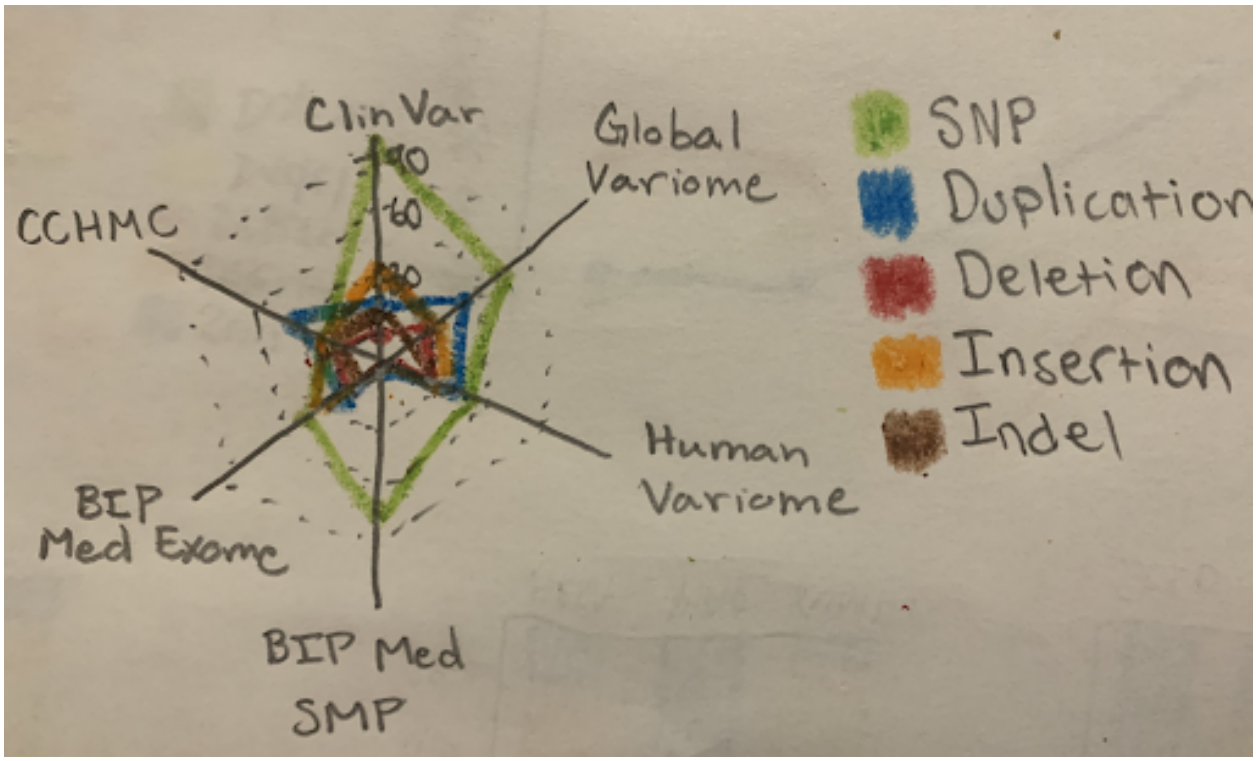


Potentially add a brush to make it so that you can zoom in.

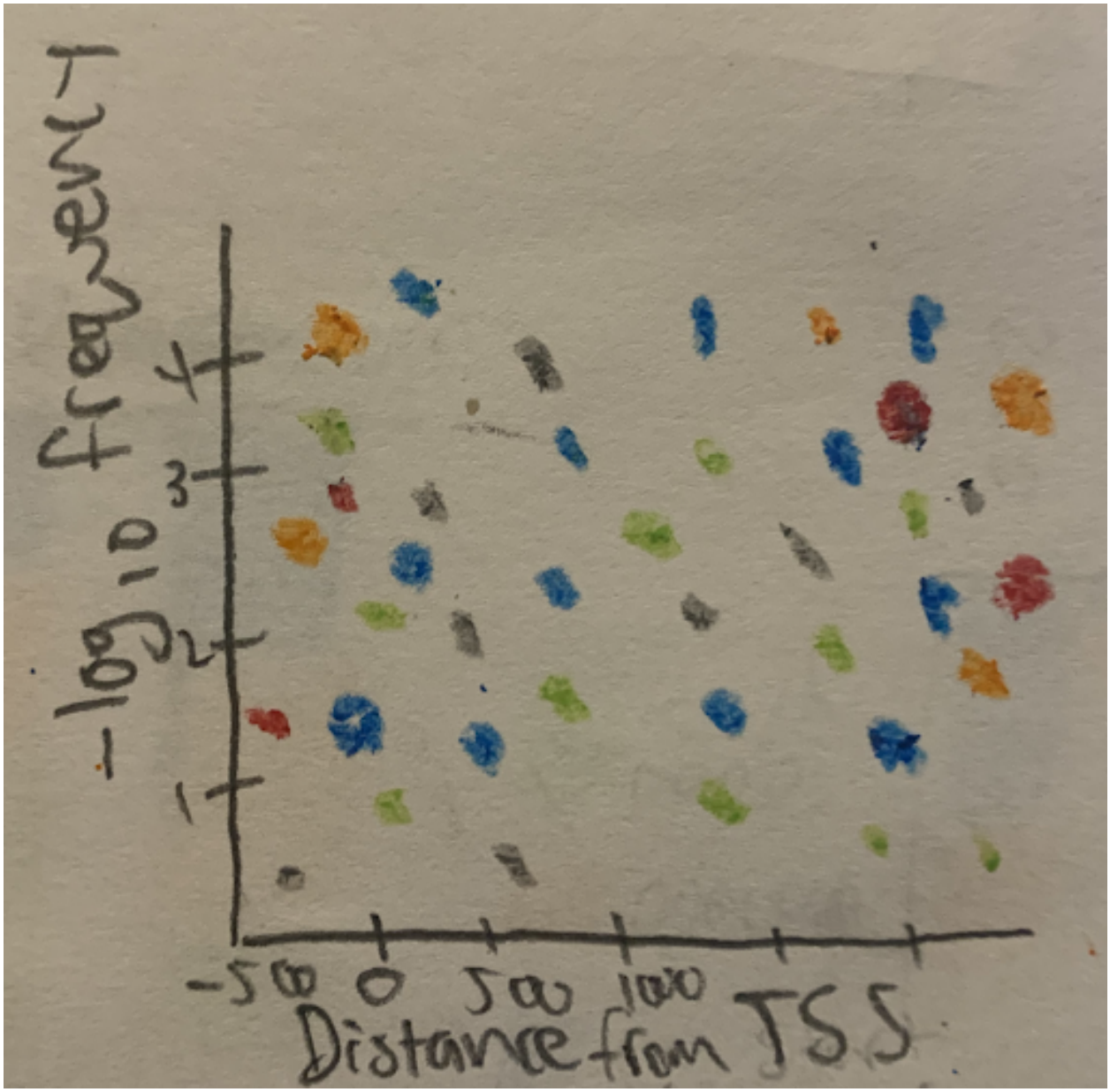


This graph could also be redrawn with the color representing the variant type (e.g. red is single nucleotide variant, green could represent duplication, etc.). Instead of using frequency of the variant, the height could represent the number of times a variant was reported or the number of stars the variant is given in ClinVar.

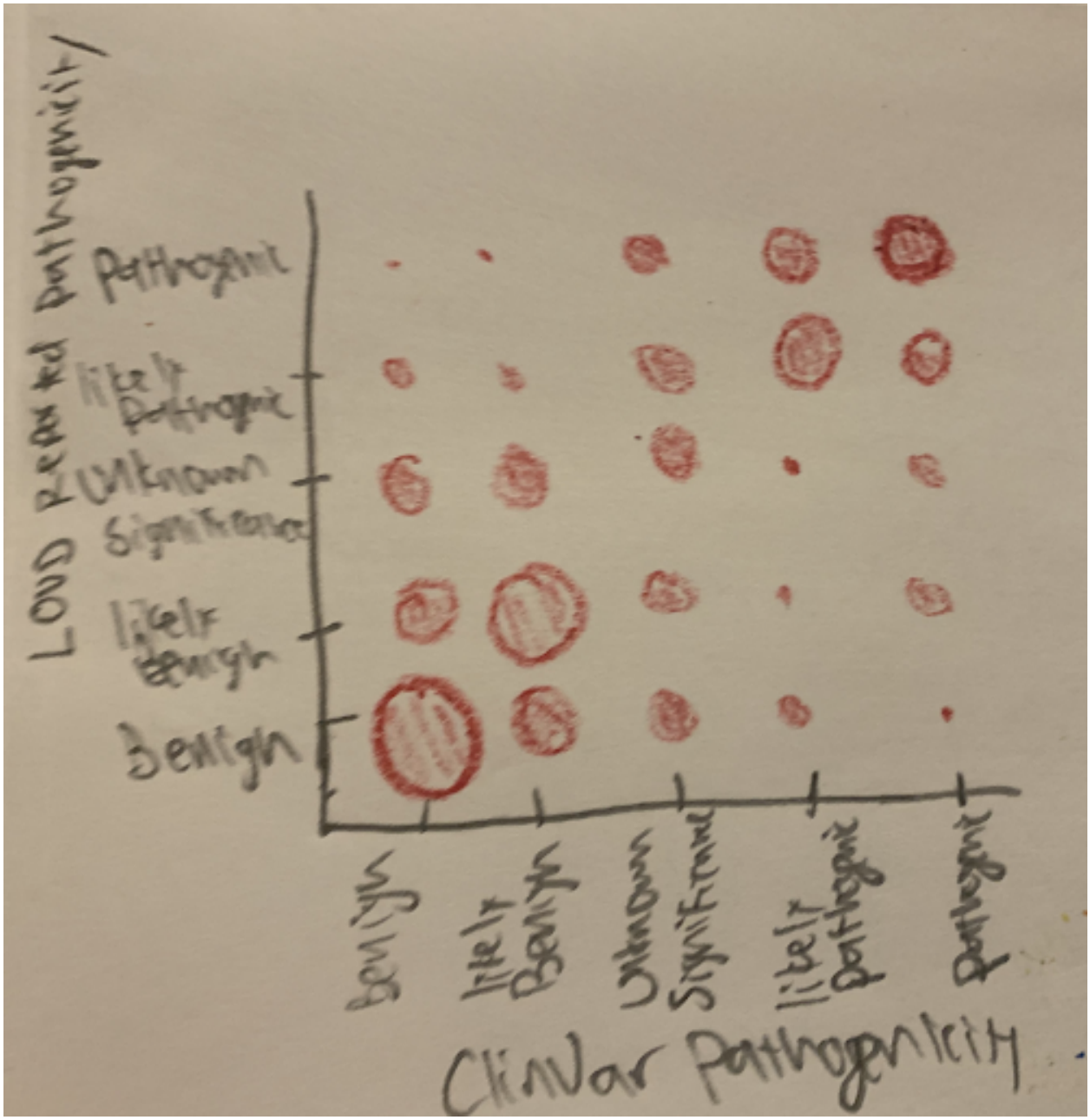
Another chart could be a radar chart that shows the number of variants of each type per database. Color would represent the type of variant, each position represents a single database and distance from center represents the number of variants of that type.



Another chart for this page could be a scatter plot. The X-axis could represent location within the gene, the Y-axis could represent the -log10 frequency of the variant. The color could represent pathogenicity (colors are same as the previous coloring), and size of the point could represent the number of times the variant was reported.

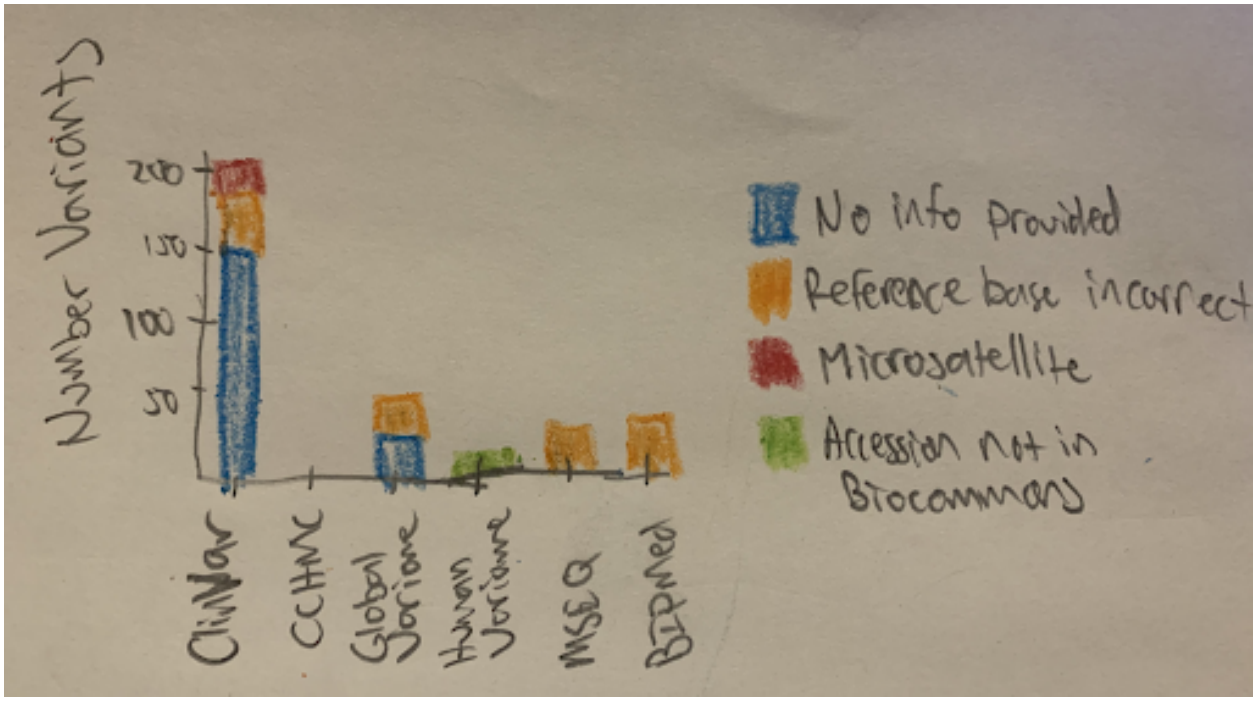


Another graph could be a scatter plot to show the comparison of reported pathogenicity of the variants. Across the x-axis could be the reported pathogenicity in ClinVar and across the y-axis could be the reported pathogenicity in LOVD. The size of the circle could represent the number of variants that fall within that category.



Or add a bar chart, maybe add the number into the middle of the circle

A final graph for the individual gene page could be a stacked bar chart showing the number of variants that failed HGVS validation per each database. The X-axis could be the database. The Y-axis could be the number of variants that failed for each category. The colors could represent the reasons that the variants failed validation. This is important to point out that a large portion of variants from ClinVar fail validation, while almost nothing from the LOVD databases fails validation.



This chart could have a button to click to switch it from number of variants that failed to percentage of variants in the database that failed for each given reason. For example, here are a couple of mutations that are valid by ClinVar standards, but cannot be normalized using the biocommons tool.
Microsatellite: NM_0001.01:c.123A[4]
Compound: NM_0001.01:c.(?-123)_(456,789-?)del

Must-Have Features

The web page must have the drop-down button to allow the user to select a disease. The web page must be able to display the bar charts showing the percentage of overlap between ClinVar and LOVD databases. It must display per disease, and must be able to display per gene when a disease is selected. The stacked bar charts showing the number of variants of each type is also necessary.

For the individual gene page, the scatter plot showing the comparison of pathogenicity between ClinVar and LOVD is necessary.

The individual gene page should also show the stacked bar chart representing the number of variants of each type.

The stacked bar chart showing the number of variants that failed HGVS validation should also be included.

Optional Features

The ability to go to a second page to view information about an individual gene is nice, but not absolutely necessary. The option to switch to count of variants instead of percent of variants is nice but not necessary. It would also be nice to have the multiple line chart showing the number of variants of each pathogenicity per gene, but this is not necessary.

The ability to filter by year, pathogenicity, database, or number of stars from ClinVar is also ideal, but not absolutely necessary.

The line graph showing the location of variants within the gene and their frequency would be really nice, but not absolutely necessary.

The radar chart would be fun and add a figure that was not made using only rectangles and lines, but is also not necessary.

The scatter plot showing the distance from TSS on the x-axis and the log10 frequency on the y-axis with color representing pathogenicity would also add another figure that is not just bar charts, but is also not absolutely necessary.

Schedule

The bash script to obtain the frequency information and the SQLite database should be built by November 13. Dave will be in charge of obtaining frequency information and setting up the SQLite database.

The HTML page should be built with spaces designated for each design by November 13. Marcus will set up the HTML page outlines.

The absolutely necessary features should be added by November 20th. If anything is not completed, these features should be top priority to be added by November 27th. Dallon, Dave and Marcus will select which features they think are in their individual skill sets and decide who will make which feature.

The optional features should be added by November 27th if there are no problems with the absolutely necessary features. Dallon and Dave will work on the remaining features.

CS 5630 / 6630 Project Peer Feedback

We were reviewed by Kaelin Hoang and Sunny Siu

Kaelin and Sunny said that the idea behind our project was really good and that it was an interesting story to tell. They asked if this data could be useful for clinicians to view. They were impressed that we had programs to gather data from so many different sources and that we were going to do this for so many genes. This is an awful lot of data.

Specific points that we went over:

General Questions

- Are the objectives interesting to the target audience?
    - Yes, the objectives will be helpful for portraying information to people working with genetic information including clinicians
- Is the scope of the project appropriate? If not, suggest improvements. Is the split between optional and must-have features appropriate? Why?
    - Yes, the scope is met, but there is a concern that this may be too much. This is a complicated topic and it may take a lot of explaining for members of this class to understand what we are doing. Additionally, this is TONS of data to be working with considering it is such a short time to get the program together. We aren't even started with coding the website, only with gathering the informatoin. There are not very many must have features, but they are all complicated because of the amount of data being portrayed.
- Is the visualization innovative? Creative? Why?
    - Yes, they thought this was a very creative way to show information about genetic variants. It is in such a format that it is easy to understand, even for individuals that have very little understanding of genetics. They especially liked the gene diagram.
- Does the visualization scale to the used dataset? Could it handle larger but similar datasets?
    - They did not comment on this, but all of the figures can scale for all possibilities in the data we are using. We do not plan to add new diseases, but if new diseases were added they certainly would not have so many genes that it would make our visualizations no longer useful. However, if

we do plan to move on to a very large gene list (such as the genes related to deafness), this would need significant adaption.

- Is the project plan detailed enough? Is a path to the final project clear?
    - Yes, the plan is very detailed, and the path to the final project is clear.
- Is an interesting story told?
    - Yes, the story is interesting because it can be used to help people determine if variants are pathogenic or not.

Visual Encoding

- Does the visualization follow the principles used in class?
    - Yes, the basic principles for design of visualization and usability and interpretability are followed.
- What is the primary visual encoding? Does it match the most important aspect of the data?
    - They specifically mentioned that they like how we have several different types of encodings. They said that they thought it was creative to have so many different ways of showing the information. They really liked that it starts off with a broad picture, then you can zoom in to look at genes within a specific disease, then you look at info about a single gene.
- What other visual variables are used? Are they effective?
    - They noticed the different views we had designed and said they looked like they would effectively allow the user to investigate multiple aspects of the data.
- Is color sensibly used? If not, suggest improvements.
    - Yes. Color scheme used was copied from python, which uses a color-blindness sensitive scheme. Any changes in color will still take into account colorblindness and make sure the colors are easily distinguishable.

Interaction and Animation

- Is the interaction meaningful? If not, suggest improvements.
    - Yes, they mentioned how they liked that each visualization allows us to drill down into each level of the data. They also thought adding a brush function would be helpful on the gene diagram. They liked the multiple layers of filters.
- If multiple views, are they coordinated? If not, would it be meaningful?

    - We do not have multiple views, unless you consider the zoomed view on a gene. Multiple views wouldn't be particularly meaningful in this case because each visualization is planned to be independent of all other visualizations.
- Is there any animation planned? Is it clear? Is it intuitive?
    - We do not have any animation planned at the moment. We will incorporate transitions into the visualization when selecting different views.

Mentor Feedback

Our progress was reviewed by Devin Lange on November 20, 2020

During the meeting we went through the current status of the website and all figures. We had proposed several figures, but drafts of only three of them were made. Probably the greatest thing about the website is that all filters were effective for all figures made. All figures include the ability to filter genes by database (7 options), variant type (5 options), reported pathogenicity (7 options) and review status (3 options).

Overlap Graphs

The first figure was the graph to show the overlap of variants between ClinVar and LOVD. The graph correctly filtered based on the selected input from the user. The graphs looked good, but did not have any labels. The labels needed to be added. Additionally, the colors were perhaps not ideal. The colors were chosed because they matched what python uses as default colors becuase they are well suited for people with colorblindness. However, using green, orange and blue does not really represent overlap well. Devin recommended changing the colors for something that would represent mixing, such as blue for ClinVar, red for LOVD and purple for the overlap between the two. Another option that would look nice if we could use both colors for the middle but have them in a striped pattern like a candy cane to represent that they are present in both.

Another thing Devin mentioned we should do was have interactivity. It already filters based upon several user specified criteria, but there are two graphs that show the same information in different ways and they should be linked. One shows the overlap and exclusivity by percentage, the other shows the same information by percentage. If you could click one section of a bar in one graph, we would like it to highlight the related section in the other bar graph.

Line Graphs

We have one figure that shows line charts indicating the proportion of variants of each given pathogenicity. Devin pointed out that linking them makes sense as they are in the order from lowest pathogenicity to highest pathogenicity, but the two at the end ('conflicting' and 'not reported') are not actually on that scale. Devin recommended we add circles at each point along the line, and then make the line stop when it got to the last two. That way it would not look like they were 'extra pathogenic' and 'super pathogenic'.

Devin also pointed out that these ones need to be more interactive. For instance, if they were lower opacity and then became dark on hover, that would be nice. If you clicked on a circle, a tooltip could show you what the number and percentage of variants for that gene for that pathogenicity are.

Again, labels were not present. They should be added.

Scatterplot

The scatterplot looked really good. Once again, labels need to be added. This apparently was not a concept on any of our minds until we finished making the graphs. This figure was not yet finished, so the dots were all red. The colors need to be updated to represent the pathogenicity as we had said we would.

It would be nice if you could hover over any given point and have a tool tip that would show the variant information.

The scatter plot shows some great information and would really be ideal for a storytelling piece. No storytelling has been added yet, so using a preset filter could show something really cool here.

Other things

Aside from the changes to the individual charts, there were a few other things Devin discussed with us. We didn't have a storytelling piece set up yet. He talked about using a predetermined set of filters that show up when you click a button, and highlight the parts of the figures that are important.

We also had a few figures that don't make sense any more because we have fewer genes and less data now. Devin had suggested that we just stick with the MLD genes instead of including all three diseases as we had originally planned.

Project Updates

As things have progressed throughout the semester, we have made a number of changes to what we plan to do for our final project. Here we will discuss the larger changes to our plans.

Limiting the Dataset

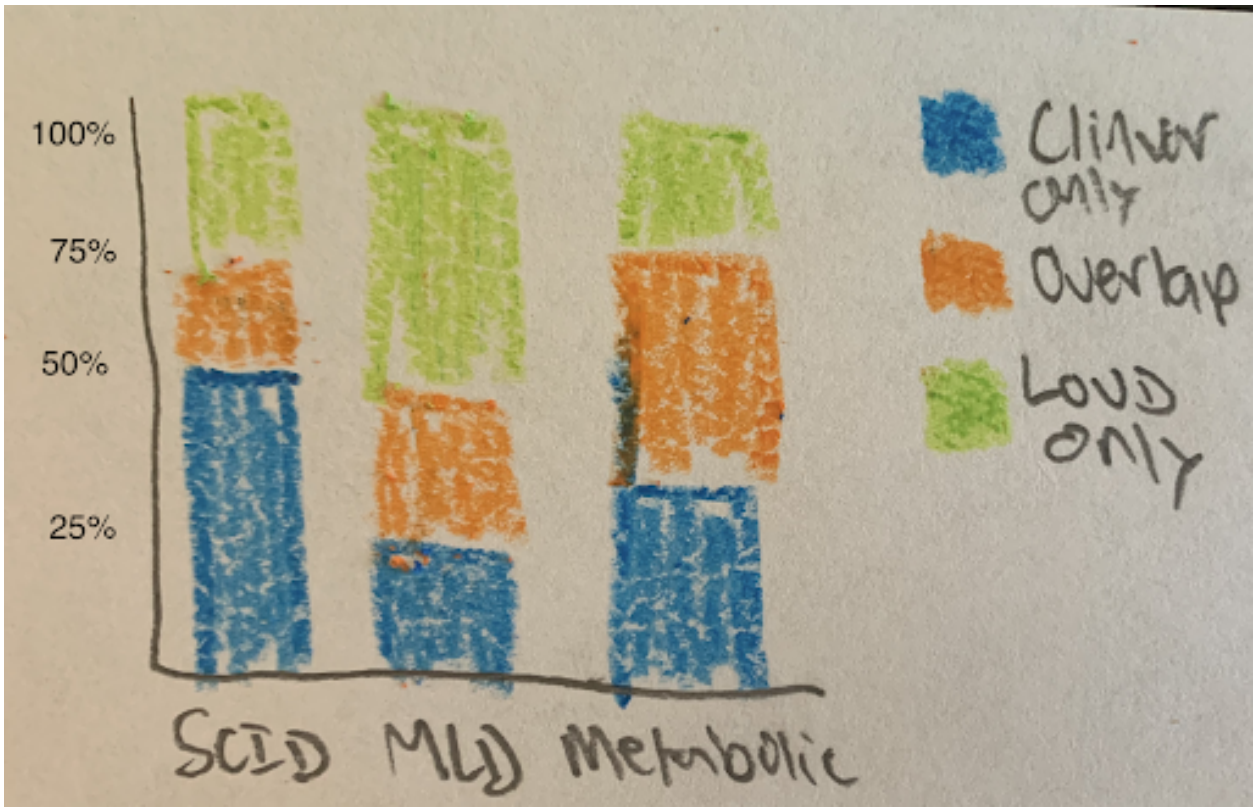
We originally proposed to use 3 different diseases with 55 genes combined. We obtained data for all 55 genes and prepared everything as if to use all 3 diseases. However, we realized that this was simply too much for us to work with. We have removed two figures that show the overlap between LOVD and ClinVar, and limited the dataset to only Metachromatic Leukodystrophy (MLD). This means that we will be working with only three genes. This will still allow us to make the majority of the figures, but will be much simpler for us to sort and work with.

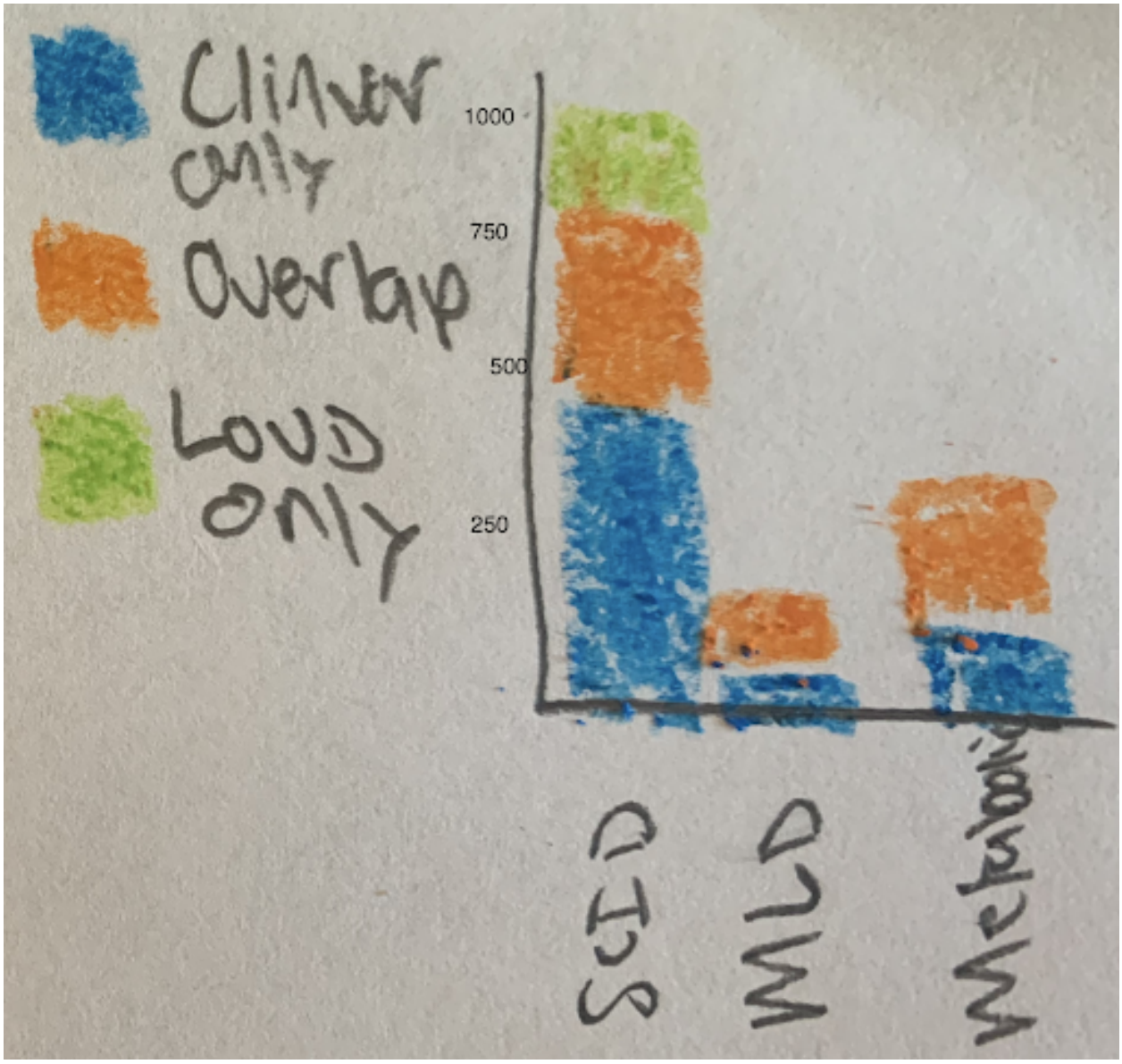
No Longer Using SQLite

We originally planned to read everything into a SQLite3 database, then host a flask server that would allow us to use SQL commands to obtain information for building the figures. Using SQL was really the only way to make the queries fast enough to reasonably update the figures. The SQLite database is actually built and present in the SQL folder. However, setting up a flask server to be able to return the queries turned out to be too difficult. The decision to limit the dataset to a single disease means that JavaScript will be fast enough to perform all of the calculations without sending queries to a SQL server. Therefore we abandoned the idea of using SQLite and are now simply using JavaScript.

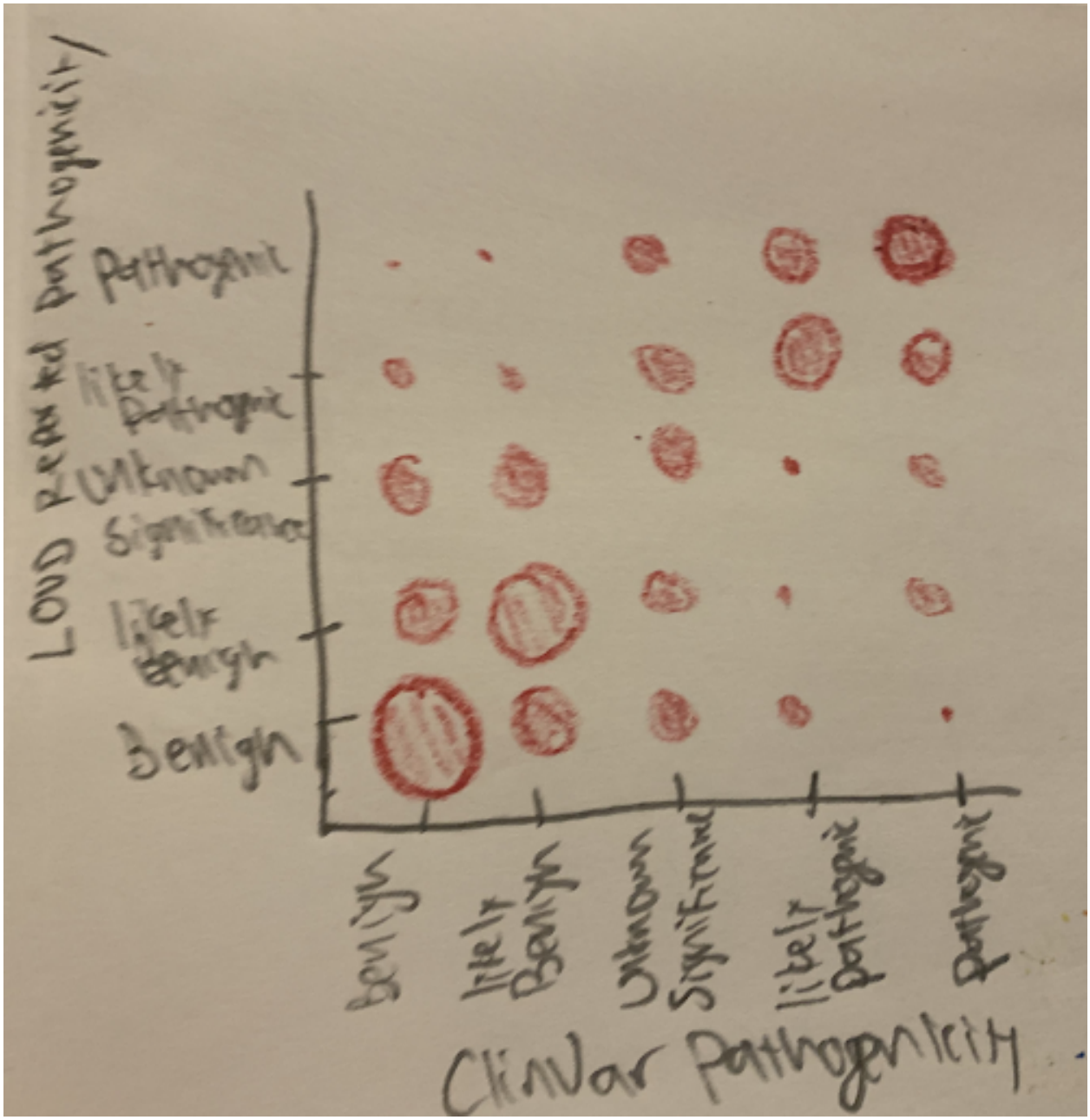
Removing Certain Figures

We originally had two figures showing the overlap between ClinVar and the LOVD databases per each disease. We now have only one disease (MLD), so these figures would not really be relevant. All of the information from these bar charts could also be found in other figures, so we have chosen to remove these two figures.

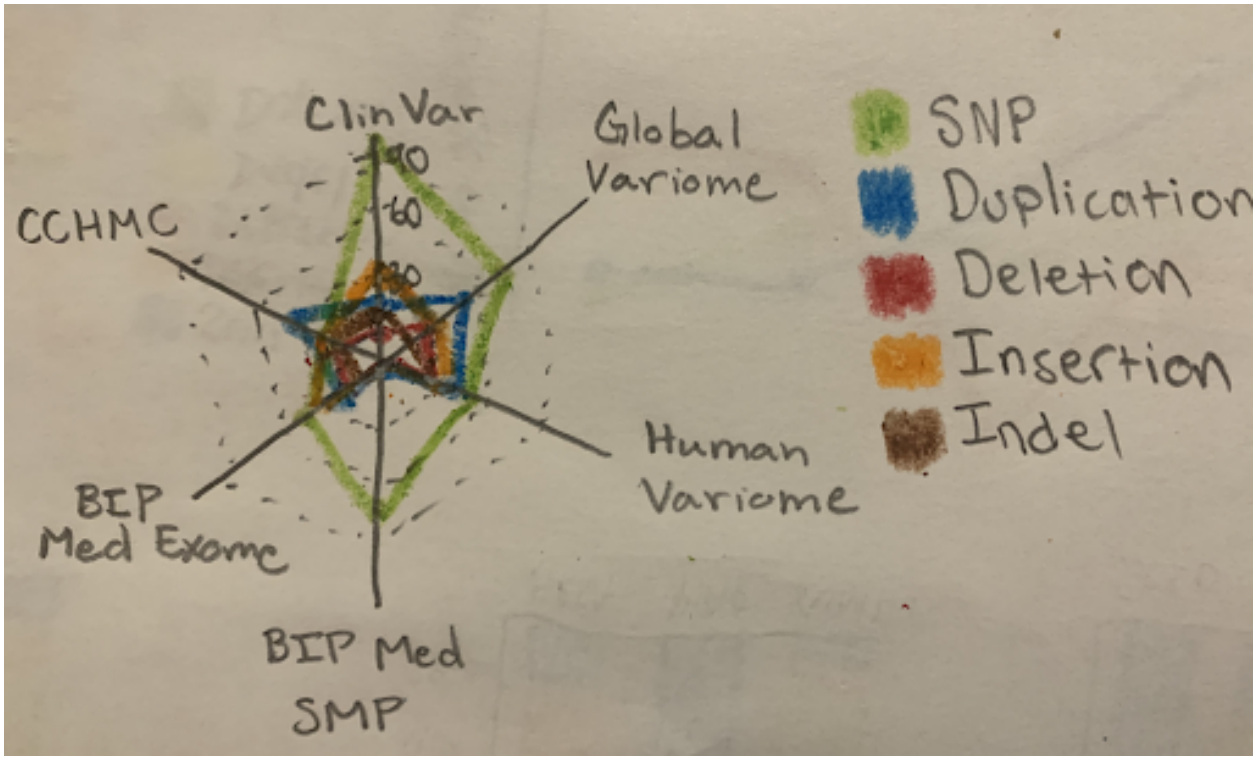




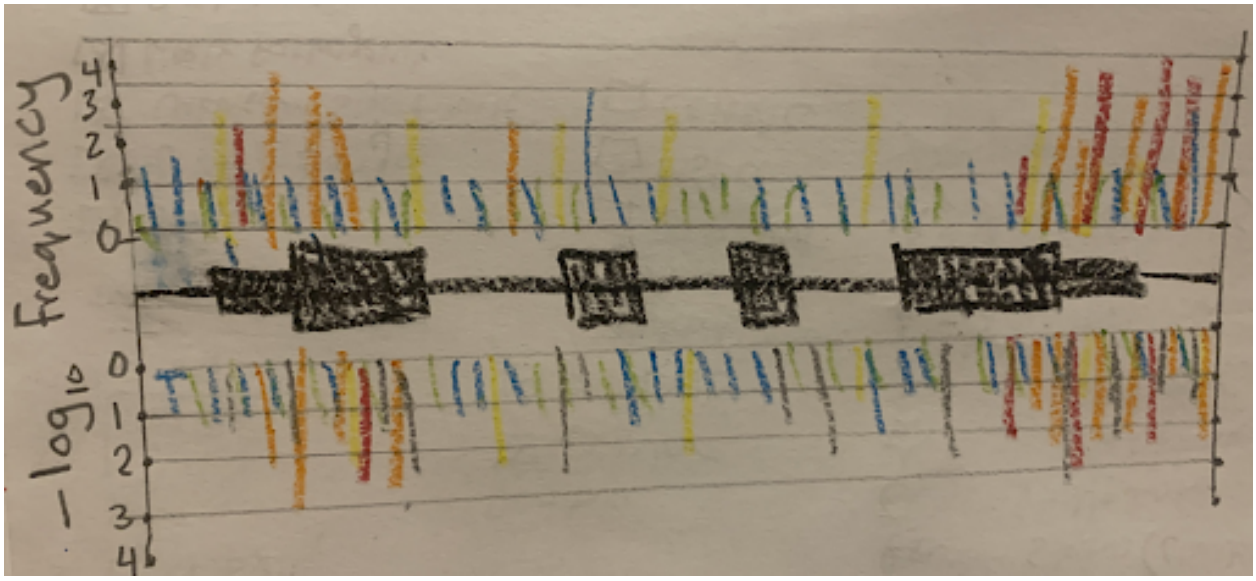
We also had planned to include a figure showing a comparison of the pathogenicity reported for variants present in both LOVD and ClinVar. Unfortunately, LOVD contains reported pathogenicity information only for some genes. As it turns out, all three of the MLD genes have no reported pathogenicity, so this figure no longer makes any sense.



Devin also suggested that we remove the radar chart as the information can be found in other figures and the chart itself is rather noisy and difficult to interpret.



One more figure was still relevant, but it was by far the most complicated figure proposed. This was the graph with lines representing individual variants, with the height representing the frequency (-log10 of the overall minor allele frequency), and rectangles representing the exons of the gene. Most of this information is present in the scatter plot with location (from transcription start site) along the X-axis and frequency (-log10) on the Y-axis, but this figure would have shown differences in reported pathogenicity between LOVD and ClinVar. As it turns out, all of the LOVD variants would have been gray so this would be far less interesting. Considering that this graph would have take so much effort to make and that the important information can also be seen in the scatter plot, we chose to remove this figure.



Removing date from the filter criteria

In addition to these figures being removed, there was one filter that was removed: the submission date. Originally we were going to have the option to filter based on the year that the variant was submitted to the website. Unfortunately, almost all of the LOVD variants and a significant proportion of the ClinVar variants were found to have no reported date. As a result, we chose to remove the option to filter by year of submission.