Lyme Disease Analyzed with GSE145974 Data

August 27, 2020

<https://public.tableau.com/profile/janis5126#!/vizhome/TreeBoxChartLymeDiseaseGSE145974/TreeBoxChartLymeDiseaseGSE145974?publish=yes>Lyme disease is a disease that infects many people annually, and is considered common due to the amount of people who get it. It is spread or caused by ticks. Symptoms can be deadly and lead to disability. This data obtained has a recently published research article on lyme disease available on [pubMed for free](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7078463/pdf/mBio.00047-20.pdf). The researchers who published this data sought out possible gene targets for identifying lyme disease early so that patients who don't know they have it can be tested by their blood mononucleated cells to see if the genes they identified in their study would identify a new person as having lyme disease so that they could get treated early and possibly with great benefits avoid the deadly symptoms of lyme disease that can attack the nervous system.

This blog isn't a summary about the resource but showing what I discovered when analyzing the data that I de-standardized after initially assuming it was log2 normalized based on the method I could get the initial values back. There is a linux version of the fastq files but I am having technical difficulties in my virtualBox Ubuntu disk image source working and also the SRAtoolkit to work. Good tutorials for my machine are hard to find, because some important steps are missing or if available are for other operating systems or software versions, etc. It is not uncommon to have a greatly written tutorial show results, but when you use your machine to run the same commands the instructions fall short. I have had this happen with many tutorials after googling them, some even from Kaggle, because they (Google) expects you to use their notebook to run code, and it doesn't always work the same. And actually, when I spent $10,000-$15000 just on the bioinformatics portion of my course, I had one professor who was not a mentor and in fact a bully who didn't want to read or help or show actual examples of bioinformatics research, but had us learn from pulling research articles online, where that is like finding a needle in a haystack. I guess we were part of the new batch of the now ubiquitous class of data scientistst to do that type of research. But examples and good mentorship and not bullying of students is key to learning and discovering and sharing. But now there are more examples, and better ones, just not for extracting the fastq files of NCBI donated GEO data. More stuff to learn on our own and 'synthesize' so that we can find out how to really get things working and not just read about some methods that worked for someone else at some other time on a different machine with different settings or software versions, etc. Fastq files are really the key, because as you get more into research and higher education, you realize how much error there really is in standard measures or values, or even accuracy in prediction. Fastq and fasta are the actual sequence maps that are billions of nucleic acids long and many versions or copy number variants with different alleles being dropped, skipped, repeated, whatever can occur. Those sequences are mapped to other known genes and if it has a certain threshold of probability of being that gene set by the scientists or program or software then it is classified as that gene. As a data scientist I don't pretend to be an absolute authoritarian expert in bioinformatics as I minored in it in my master's program. Just in interpreting the data, because diseases can happen at any point in epigenetics of the non innate gene expressions at the replication, at the exon an intron levels, at the network signaling for the number of gene copies to produce, at the amino acid stage of protein building, at the structural building phase, etc. Many factors from environmental or internal stress create those changes and how our bodies adapt and microscopically solve things like infections or growth, or working out, or constant mental or physical stress on the body. That is what is important, and you can have a bioinformatics talk to you and tutor you and pay for their insight and still not get what is known generally. It comes down to objective deductive reasoning on each scientist you talk to and how they interpret the findings. The process doesn't even sound that complicated, and their are literally billions of different versions of genes that the world produces in all mammals, and tens of thousands that are known to some extent. What I am seeking to find is those genes that are being obviously over expressed or suppressed dramatically in the disease or treated samples when compared to the healthy samples. Because there is information there, then once you get those genes from data values lab scientists have provided and tested and retested for accuracy, then you send your results to those scientists that look at the motifs and copy number changes from the sequence files and find those genes that were classified as such and pull all those copies to then test if there is a high probability of having a pathogenesis or disease causing relationship with those genes and specifically which alleles or genotype versions of the gene's various copy numbers are most likely contributing to a disease or treatment outcome response. So, keep that in mind as you scroll through this data.

This blog looks at lyme disease through a GEO series study, GSE145974, that analyzed patients with lyme disease over two years. Patients came and went during the trial study with few dropping out or not being there for the six month blood extraction of peripheral blood mononucleic cells (PBMC) blood samples after antibiotics. There were healthy samples totalling 21, patients with acute lyme disease infection totalling 28 samples, one month after antibiotics there are 27 patients who gave their blood, and six months after antibiotics treatment there were only 10 patients who gave their blood. More details are in the NCBI gene summary and also on the published research article that donated these samples. The processing was done using Affymetrix and the human genome 37 data using expression profiling by array. The ages and genders weren't given as identifiers for these samples and so can't be compared like a previous analysis on COVID-19 using GSE152418 data. The data that was readily available did have all the platform gene identifiers which was great, but was already log2 normalized. There are various methods for log2 normalizing and I used a typical one of subtracting the min of each vector then dividing by the max-min of same vector, then taking the log base 2 of the results for the output values to be scaled between 0 and 1. The rpubs document highlights and tracks all the steps to these results. The charts were then made to see visually any results by class and gene. All images are links to the Tableau chart and can be downloaded by clicking the upper right down arrow for downloading the chart and underlying data table. But you have to have the Tableau Public Server that is free downloaded and installed on your personal computer for it to work. You can also use your work computer if you are allowed. The rpubs document is available at: <https://rpubs.com/janisharris/lymeDiseaeGSE145974>. And what files were the right size to upload into github are located at: <https://github.com/JanJanJan2018/LymeDiseaseGSE145974>. Otherwise, if the files are too large, the original NCBI access link will provide the underlying original data, and you can use the Rmarkdown file to run the chunk by chunk commands to create those other supplemented tables. There are 46,000+ genes and after aggregating by grouping by genes on mean values for each gene, it was reduced to 19,000+ genes.

The Rpubs document of mine has the links to NCBI:

"This data analysis is on lyme disease using GEO series data made readily available in its normalized state from [GSE145974](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145974) on ncbi.nlm.nih.gov as the accession number. The data is from the [platform GPL13667](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13667) and the [series data](https://ftp.ncbi.nlm.nih.gov/geo/series/GSE145nnn/GSE145974/matrix/). There are also some CEL/TAR files for Ubuntu but I couldn’t get my ubuntu machines to recognize it, and the instructions and tutorials for accessing the SRAtoolkit and using the Windows Ubuntu app, didn’t avail, so I am using the text files only.If you have a windows 10 tutorial on running SRAtoolkit using the ubuntu app for windows or getting the cel files to work on ubuntu with a VirtualBox disk image of ubuntu that works because you tried it within 24 hours and it worked exactly as explained, please share. I have yet to get those up and running. Possibly the new updates to virtualBox or my other apps, like docker or MongoDB or Tableau are interfering. I am not going to waste time figuring it out, its a time trap.

All the data was there as far as being filled out with values for the feature names, because I do recall exploring some platforms and series downloadable text files and only the header information was there and none of the values. The method used for processing was expression profiling by microarray on peripheral blood mononucleated cells (PBMC). The values seem to be scaled or normalized already as the values are inclusive of negative values."

Here is the data series research article information:

"Global Transcriptome Analysis Identifies a Diagnostic

Signature for Early Disseminated Lyme Disease and Its Resolution" authored by the following researchers:

Mary M. Petzke,a Konstantin Volyanskyy,b Yong Mao,b Byron Arevalo,a Raphael Zohn,a Johanna Quituisaca,a Gary P. Wormser,c Nevenka Dimitrova,b Ira Schwartza

They are with the Department of Microbiology and Immunology, School of Medicine, New York Medical College, Valhalla, New York, USA bPhillips Research North America, Valhalla, New York, USA

Division of Infectious Diseases, Department of Medicine, New York Medical College, Valhalla, New York, USA

\*\*Citation\*\* Petzke MM, Volyanskyy K, Mao Y,

Arevalo B, Zohn R, Quituisaca J, Wormser GP, Dimitrova N, Schwartz I. 2020. Global transcriptome analysis identifies a diagnostic signature for early disseminated Lyme disease and its resolution. mBio 11:e00047-20. https:// doi.org/10.1128/mBio.00047-20.

Editor Steven J. Norris, McGovern Medical School

Also, as a side note, I did have an inlaw die of Lyme Disease and he was in his 50s. Not an inlaw I ever met, but my sister's stepson's half-brother's biological father who lives in Oregon. Oregon is a very green state and many hikers do get ticks and subsequently or consequently lyme disease. I have heard from a separate article I don't recall as it was weeks ago, that it is spread within 24 hours of being bitten. And ticks latch onto people and pets. Many don't know they have them and have to check. It might be from numerous bug bites like mosquitos that they don't notice, or a neuro toxin they possibly release like leeches that prevents the host from feeling them biting them.

Lets get to it, shall we?

<https://public.tableau.com/profile/janis5126#!/vizhome/LymeDiseaseDashboardGSE145974/LymeDiseaseDashboardGSE145974?publish=yes>

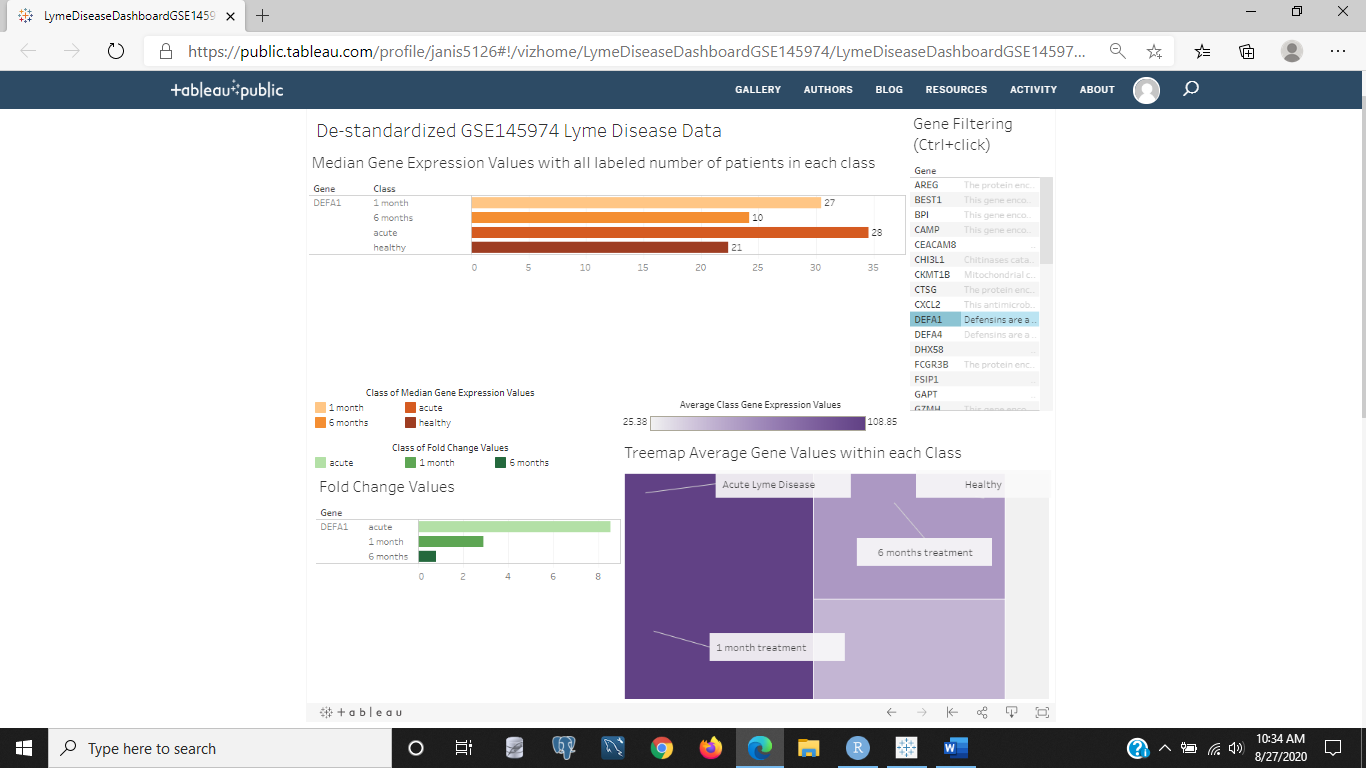


Figure 1: Dashboard showing Lyme Disease data with de-normalized data from a GEO series study, GSE145974, recently published in April of this year 2020. You can see the genes to the right with the gene summaries if you hover over the text to the right of the dashboard in the 'Gene Filtering' box. It will select only the genes you select to show the median gene expression values within each class of healthy, acute lyme disease, one month after antibiotics treatment, and six months after antibiotics treatement, with varying class sizes due to changes in patient participation and methods during the study. The top chart of the warm colors is for the median gene expression values for each gene of 43 genes that were filtered from 19,000 genes as having the most or least fold change in disease or treatment to healthy ratios for all three classes with duplicates removed from the top 10 or bottom 10 genes in each class by fold change. The lower left chart with the greens is the fold change values for each gene within each class of acute lyme disease, one month of treatment, or six months of treatment compared to healthy samples by mean values of all samples in each class. The lower right chart of the purple colors is a tree map that is categorized by class and within each class each box is a gene with the average gene expression value within that class for that gene. The upper right box shows that the gene DEFA1 was selected and it is displayed in all three accompanying charts on the dashboard.

<https://public.tableau.com/profile/janis5126#!/vizhome/medianClassValuesLymeDiseaseGSE145974/medianClassValuesLymeDiseaseGSE145974?publish=yes>

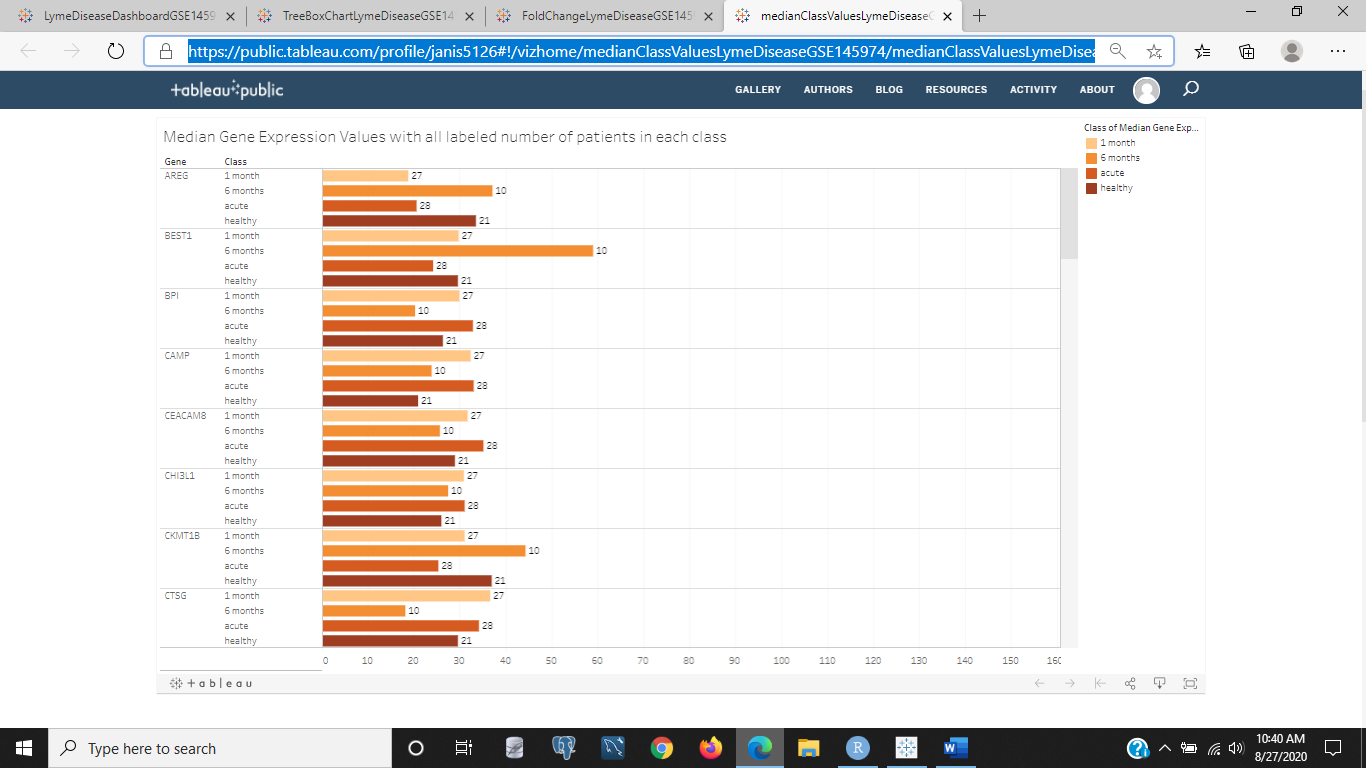


Figure 2: This is a bar chart of the median gene expression values with all labeled number of patients in each class on each bar. The median value is great because it aligns all samples within each class in order and takes the middle value as the median value or the 50th percentile. And this is a real value, not an averaged value from the sum of all samples divided by the number of samples. So, you could get a more realistic value for the bunch. But the median can also be skewed if there aren't enough samples or missing values are included or not a close distribution in values. For example, you have 10 numbers that range from 1 to 100, 6 of them are under 30, two are over 40, and one is 99 and one is 100. The median would take 30 as the median because it is the middle value. So the 40% of the set that were much greater than 30 are considered outliers of a sort. There are 43 genes and the classes above are for the healthy class which has 21 samples, the acute lyme disease class or infected class with 28 samples, the class of patients who have received antibiotic treatment for one month with 27 samples, and the class of patients who have received antibiotic treatments for six months with 10 samples. Scrolling through the data table above, if you click the image link to the Tableau chart, CXCL2,THBS1,OLR1,OR2B11,MUC12,IL1B, and KIAA1245 are very up regulated in patients after six months of treatment compared to any other class. There are genes that are also much more down regulated after six months of treatment compared to the acute and healthy classes, KIR2DS1, HBG1, GZMH, FCGR3B, and more. The gene summaries for these genes can be found when you hover over the gene filter in the dashboard linked through Figure 1.

<https://public.tableau.com/profile/janis5126#!/vizhome/FoldChangeLymeDiseaseGSE145974/FoldChangeLymeDiseaseGSE145974?publish=yes>

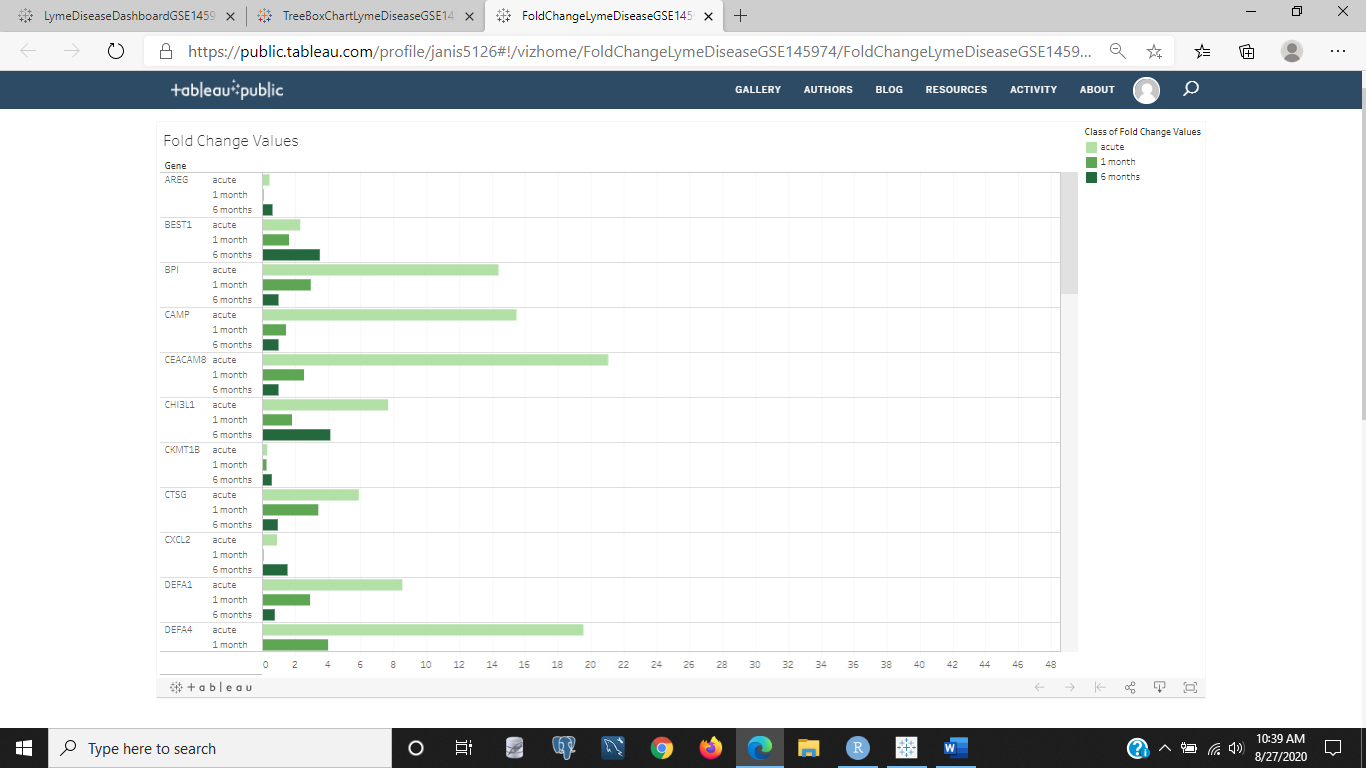


Figure 3: The above image is a bar chart showing the fold change values within each class as the class samples' mean values as a ratio to the healthy samples' mean values. There was no need to modify the fold change as was done in the COVID-19 analysis, because the values were all numeric and there were no 0s that could lead to numeric errors when dividing or inaccuracies in changes made in comparing the diseased or treatment samples to the healthy samples. If you scroll through the chart the above image links to, many of the acute/healthy fold change values are large compared to the other treatment groups, like LCN2 and LTF, and a few genes, XLST, FSIP1, and OLR1, are highly elevated in the patients who received six months of treatment compared to the acute and early treatment patients. The acute stage of lyme disease should have some genes with relatively lower than normal gene expression values compared to the treated stages meaning it is a down regulated gene. CKMT1B, IL1B, MUC12, TSIX are a few of those down regulated genes in the disease state compared to the class of patients after six months of treatment. But any gene in the acute state where the fold change is less than 1 is down regulated as a note. Hovering on each bar will display the fold change value.

<https://public.tableau.com/profile/janis5126#!/vizhome/TreeBoxChartLymeDiseaseGSE145974/TreeBoxChartLymeDiseaseGSE145974?publish=yes>

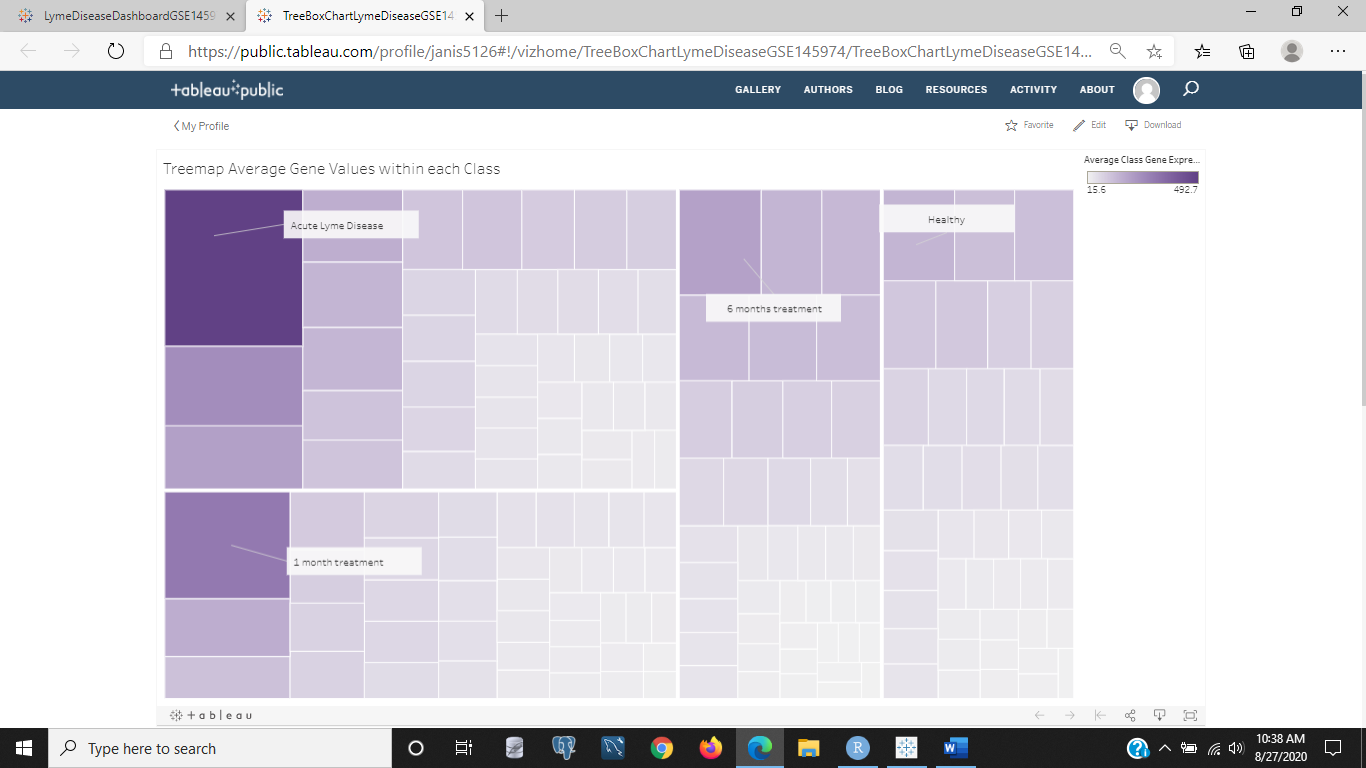


Figure 4: The above image is a treemap chart that displays the box chart type classification of our four classes and gene expression values on average for each gene within each class. The darker boxes are genes with the most gene expression and the whiter colors are those genes with not as much gene expression relative to all gene expression values. The top left corner is the acute class, the lower left is the patients after one month of treatment class, the middle right is the patients after six months of treatment class, and the far right column is the healthy class of patients. Those are sections similar to a pie chart, but as a treemap box schematic. If you hover on any littler box within a section of the tree map, it will show you the average gene expression value for that gene in all samples within that class. The darker genes are in the acute and 1 month class and hovering over those two boxes show it is the same gene, FCGR3B. And one of the lighter genes are in the center or the lower right of the treemap. POLR21 is a very light gene in the 6 month class at the lower right of that section with an average gene expression value of 15.6. This is a de-standardized value of the original log2 normalized values. To get that value, I took the original value in each sample for all genes as elements of each sample, raised 2 by each element, then I multiplied the max-min of each sample, and finally added the min of the sample. To do this by each gene probably would have made more sense, but either way it is a way of scaling the values to eliminate the negative values in the original log2 normalized data. And there were 40,000+ genes. R and many other programs like Python or SQL I am sure would fail at such a wide data table as data is better if filled with much more observations or rows than with more features than rows, which leads to the curse of dimensionality. And could probably be fixed with Principal Component Analysis (PCA), but why? We can just get the fold change values and select the ones that change the most in either direction from the initial state. And principal component analysis, shifts the data orthogonally for every feature or component and takes those features with the most variation which is exactly what we're doing. We're just rescaling it and taking those genes with the most change. Underneath, it is still the same units for each gene and sample, just adjusted within each sample. I will add more to this analysis with the original data, whose genes as target genes to identifying lyme disease are completely different than these genes, but they could both have information that predicts with 96-97% accuracy the class when training our machine learning models later.

The original data with log2 normalized values including negative values was used to make the same or similar dashboard of charts as was done above with the de-standardized lyme disease data of GSE145974. However, because of the negative values, the treemap chart was replaced with a highlight chart as the treemap removed 45 negative values. This data took a total of 32 genes (technically 33 but one didn't have a gene summary so it was omitted), and compared the most or least expressed genes as up and down regulated in the disease (acute) or treatment (1 month or 6 months of antibiotics) to the healthy samples all as the fold change of class sample mean values per gene.

<https://public.tableau.com/profile/janis5126#!/vizhome/dashboardOriginalGSE145974/dashboardOriginalGSE145974?publish=yes>

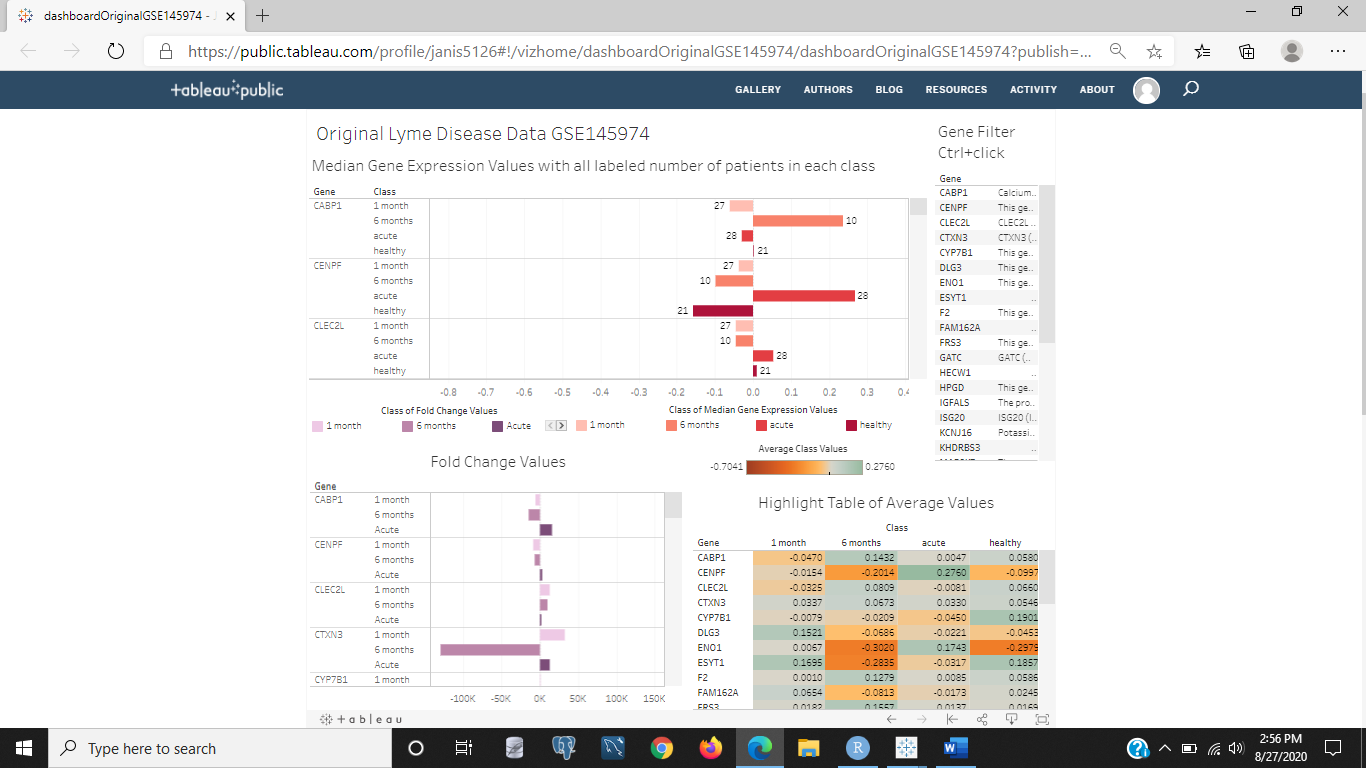


Figure 5: The above image is a dashboard of the original log2 normalized data that has negative values as well as positive values. The filter at the top right will display only the gene or genes selected if you select one then use ctrl + click on each additional gene in that 'Gene Filter' box. The other three charts will show the respective gene as it relates in the top for median gene expression values within each class and the number of samples in each class, in the lower left the fold change values for that gene in the three classes of diseased or treatment to healthy as a ratio of means, and in the lower right the highlight table with a color gradient bar of oranges and grays for the gene or genes selected. The oranges are the lowest values, or down-regulated negative gene expression values from the median, and the grays are the genes with more positive gene expression values or up regulation in each class of acute, 1 month of treatment, 6 months of treatment, or healthy. The image above will link to this dashboard to try out the different genes associated with lyme disease and treatment. I just [read an article](https://www.msn.com/en-sg/entertainment/celebrity/bella-hadid-reveals-the-truth-about-her-lyme-disease-struggle/ar-BB18oyDa?ocid=uxbndlbing) about Gigi Hadid having lyme disease since age 14 and is now 21 years old with brain fog, joint pain and stiffness, light sensitivity, headaches, anxiety, and possibly other symptoms like face paralysis related to lyme disease.

<https://public.tableau.com/profile/janis5126#!/vizhome/HighlightTableOriginalGSE145974/HighlightTableOriginalGSE145974?publish=yes>

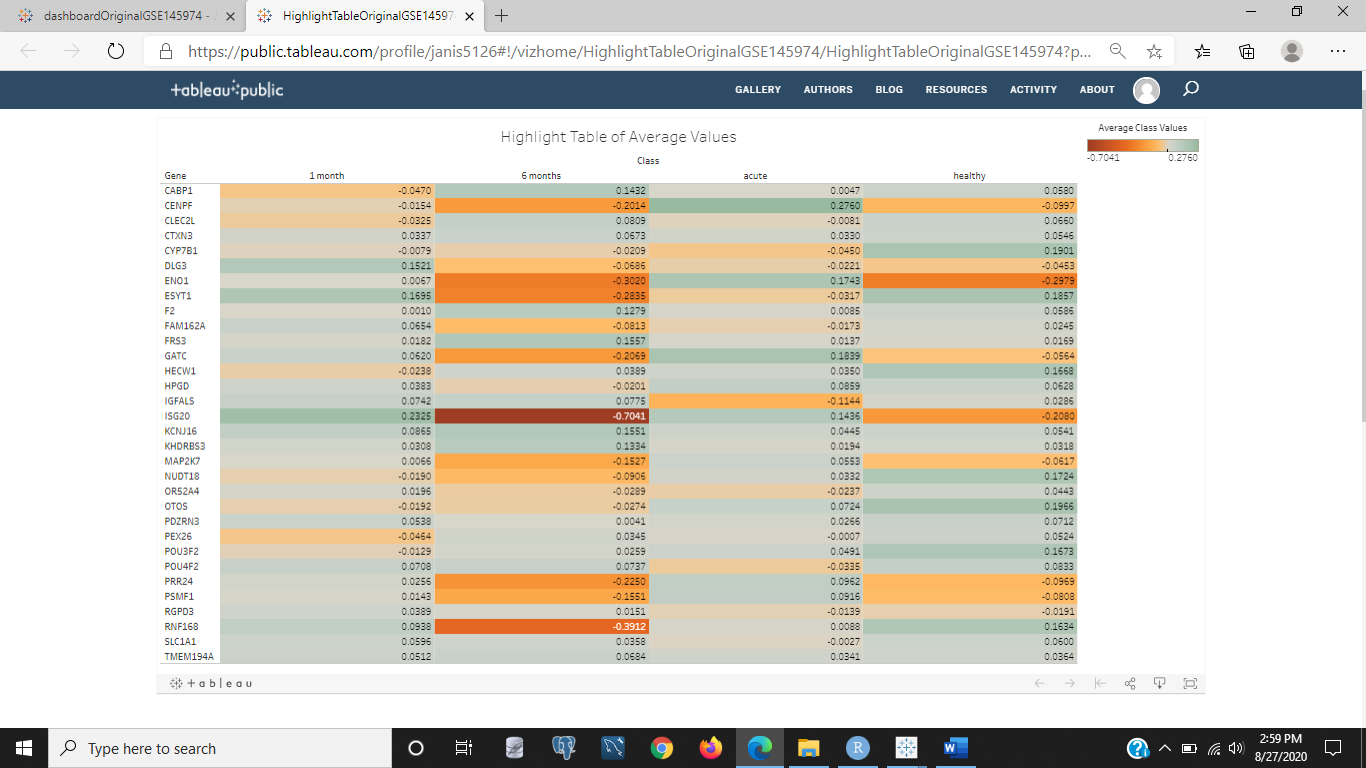


Figure 6: The above image is a highlight chart of the average gene expression values within each class of acute, one month of treatment, six months of treatment, and healthy classes. This chart was used instead of the treemap chart used in the de-standardized data because it allows negative values, and treemap charts do not and would have eliminated 45 gene values across these samples. The genes are gradient color coded, so that under expressed genes in samples having negative values are reddish-orange, and those genes with more gene expression values or up regulated are gray. Colors in between the reddish-orange and gray are for genes that didn't change as much up or down in gene expression. After six months of treatments one gene is highly under expressed, ISG20, and it is in the middle of the chart in a reddish color that indicates it is the most under expressed gene or at the end of the lowest values for gene expression. The gene summary for this gene is in the dashboard that Figure 5 links to. The Entrez gene summary says Hepatitis C and Yellow fever are associated with abnormalities in this gene and its network involvements include the innate immune system. A gene with the highest up regulation is CENPF in the acute phase it is highly up regulated. The gene summary for this gene says it could possibly have some involvement with chromosome segregation during mitosis and also that it encodes a protein associated with the centromere-kinetochore complex. Also, autoantibodies in cancer patients have been found that target this gene, CENPF. And a quick online [wikipedia](http://en.wikipedia.org/wiki/Autoantibody) search says autoantibodies are the antibodies your own body produces to attack your own body's proteins.

<https://public.tableau.com/profile/janis5126#!/vizhome/foldChangeOriginalGSE145974/foldChangeOriginalGSE145974?publish=yes>



Figure 7: The image above is to the chart of original lyme disease data fold change values for each gene across all samples. It is bidirectional as is the other charts in direction or color gradient, because this data has negative values accomodating the log2 normalized data. Negative values indicate down regulation and positive values indicate up regulation or you can think suppression versus explosion if magnitude dramatic enough relative to the neighboring genes. All these genes are the most or least expressed of all genes in the data using the original values, so they should have some change visible. The gene CTXN3 is shown to be highly down regulated with -128,817 in fold change comparison of this gene in patients' average gene expression after six months of treatment compared to the healthy samples. It was down regulated on average 10^-6 approximately less than healthy samples. That is a very large magnitude and could possibly be a target gene for having the disease or antibodies. It is important to get the treatment early to avoid symptoms, but some people still have symptoms and treatment might not work well or at all is what this could be indicating. Because the other classes of acute and one month of treatment as well as healthy don't have this magnitude of down regulation at all. Keep in mind the patients in this class was nearly a third of the original sample size 10/28 of acute patients. The gene summary for this gene, CTXN3, in the dashboard says it is a protein encoding gene that [Autotopagnosia](https://medical-dictionary.thefreedictionary.com/autotopagnosia) and [Clear Cell Adinoma](http://www.cdc.gov/DES/hcp/information/daughters/risks_daughters.html) are diseases associated to CTXN3. Autotopagnosia is the inability for one to identify his or her body parts or locate them on his or her body. And Clear Cell Adinoma is a vaginal/cervical cancer that is rare and usually diethylstilbestrol (DES) exposure in utero of a female's mother. The daughters of moms exposed to DES are more likely to get Clear Cell Adinoma and a gene that is highly underexpressed in our 6 months of treatment group, CTXN3, is associated to that disease. Either by having lower risk by not producing as much, or increased risk due to not producing much of it as the healthy and acute disease phase are.

<https://public.tableau.com/profile/janis5126#!/vizhome/medianOriginalGSE145974/medianOriginalGSE145974?publish=yes>

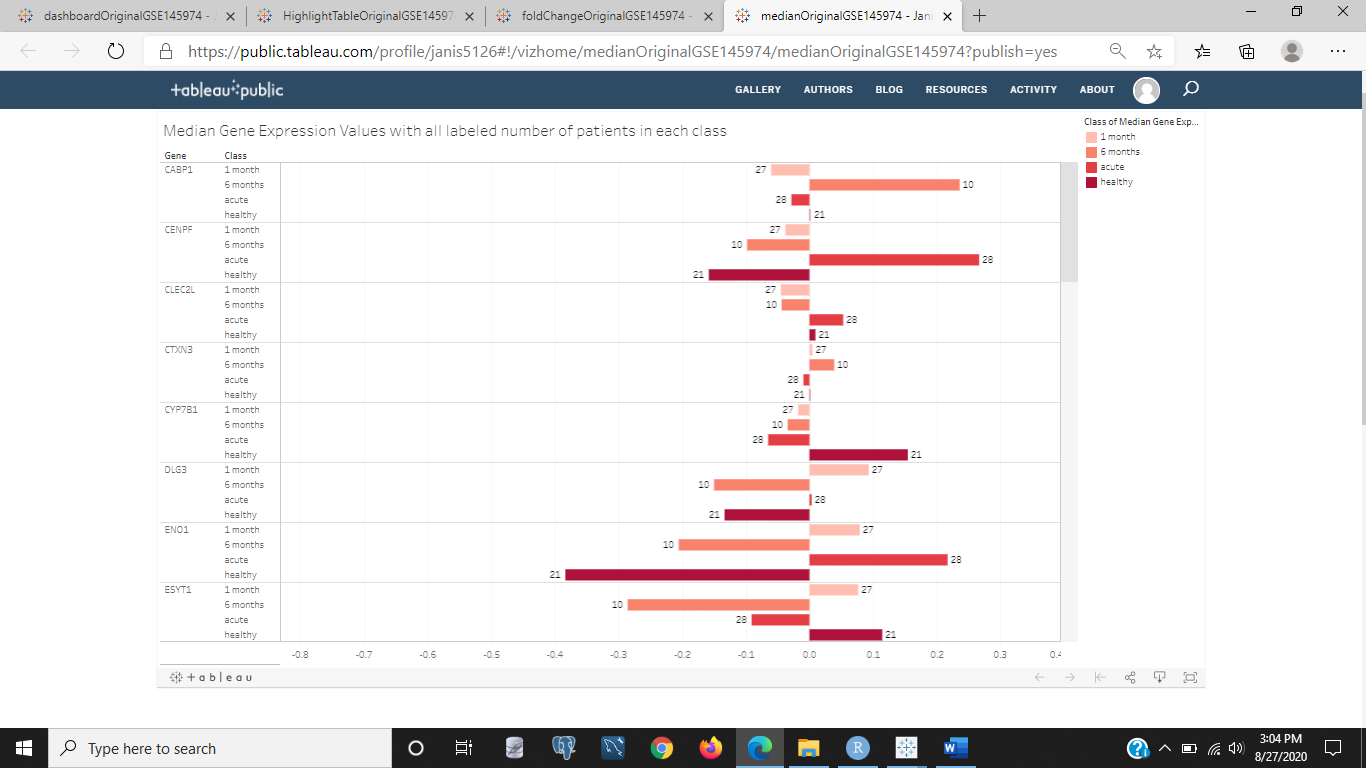


Figure 8: The above image is to the bar chart that is bidirectional like Figure 7 of fold change values, but this chart is of the median gene expression values across all four samples for each gene. The number of samples in each class is also labeled on each bar. Scrolling through the genes in the chart you will see other genes like ENO1 which has a gene summary stating it encodes alpha-enolase, one of three enolase isoenzymes found in mammals. This gene is associated with an autoantigen in Hoshimoto encephalopothy, another autoimmune contributor it sounds like. We see it is dramatically under regulated in the healthy samples and also under regulated in the samples who received six months of antibiotic treatment. But in the acute phase it is up regulated almost 50% more than the healthy samples and in the acute samples it is also up regulated but by about 25% of the healthy samples. ISG20 is very highly under regulated in the 6 month class at about 10 fold the amount of the healthy class median values which is also under regulated. We saw this gene earlier in our highlight chart as being associated with yellow fever and hepatitis C as well as innate immunity network signaling. It is the most under regulated gene in all. A gene, RNF168, is also highly under regulated in the 6 month class, but the healthy class and 1 month class are up regulated in this gene by 3-4 fold more than the 6 month class by visual inspection. This gene, RNF168, has an Entrez gene summary that states it is involved in DNA Double-Strand Break (DSB) repair, and that it has mutations associated with [Riddle syndrome](https://en.wikipedia.org/wiki/RIDDLE_syndrome). Wikipedia says this is a rare genetic disease that causes radiosensitivity, ImmunoDeficiency Dysmorphic features, and learning difficulties as an acronym meaning.

There are a lot of different genes with useful information and they are the top genes in changes in gene expression in either data, but we still need to test these genes to see how they compare using machine learning to see how well the classifications can be predicted by these genes. We will get to that later but soon.

<https://public.tableau.com/profile/janis5126#!/vizhome/genesAcrossEachSampleGSE145974/genesAcrossEachSampleGSE145974?publish=yes>

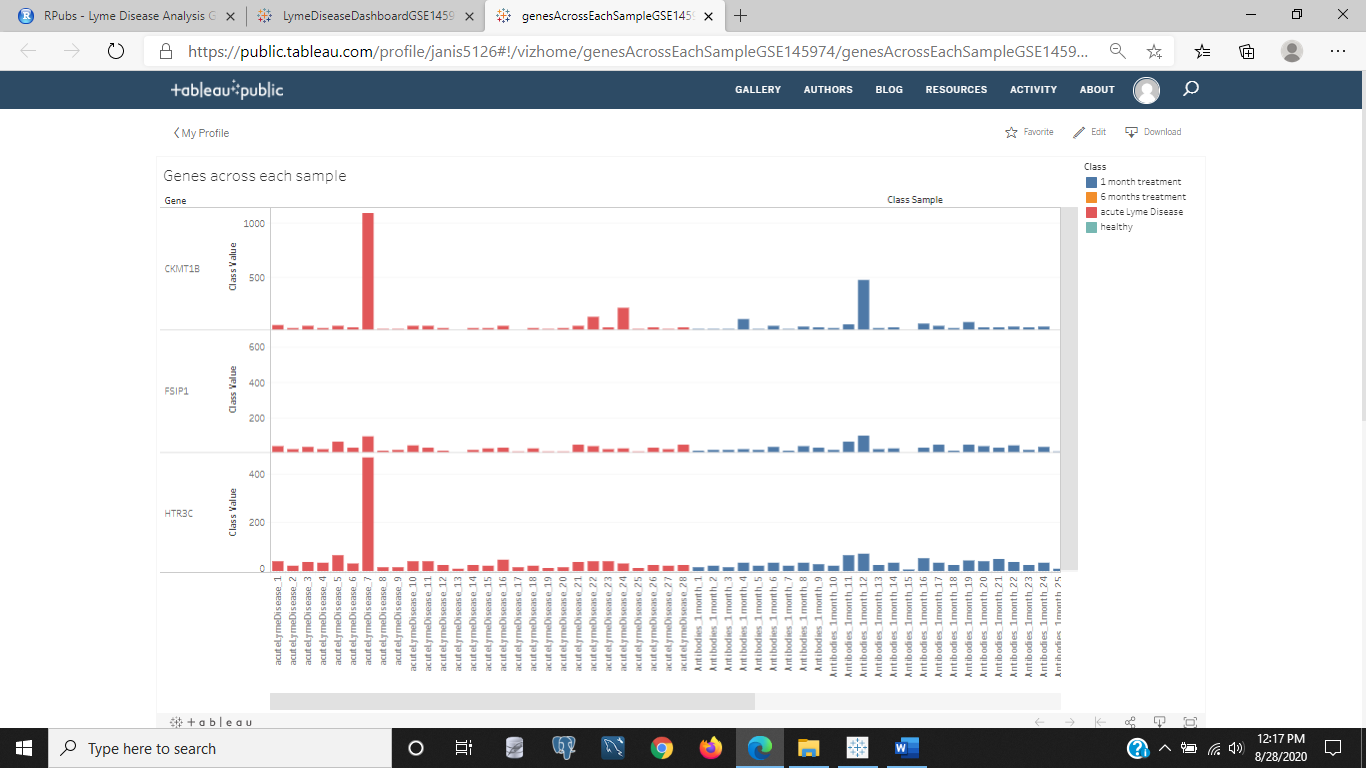


Figure 9: I added the sample chart to see the groups of individual samples within each class of healthy, acute lyme disease, one month of antibiotic treatment, and six months of antibiotics treatments, after realizing, some of the genes' fold change values are skewing the data greatly. And we can see in the above image of this chart (linked to through the image) that these samples in this set of genes that skewed our data when running some machine learning algorithms were samples: 7 of the acute class samples, sample 12 of the 1 month class samples, sample 10 of the 6 month class samples, and samples 1, 11, and 12 of the healthy class samples. I want to remove these samples and run some machine learning on the set, or just take the median sample values instead when deriving the fold change values.

<https://public.tableau.com/profile/janis5126#!/vizhome/dashboardMedDropGSE145974/dashboardMedDropGSE145974?publish=yes>

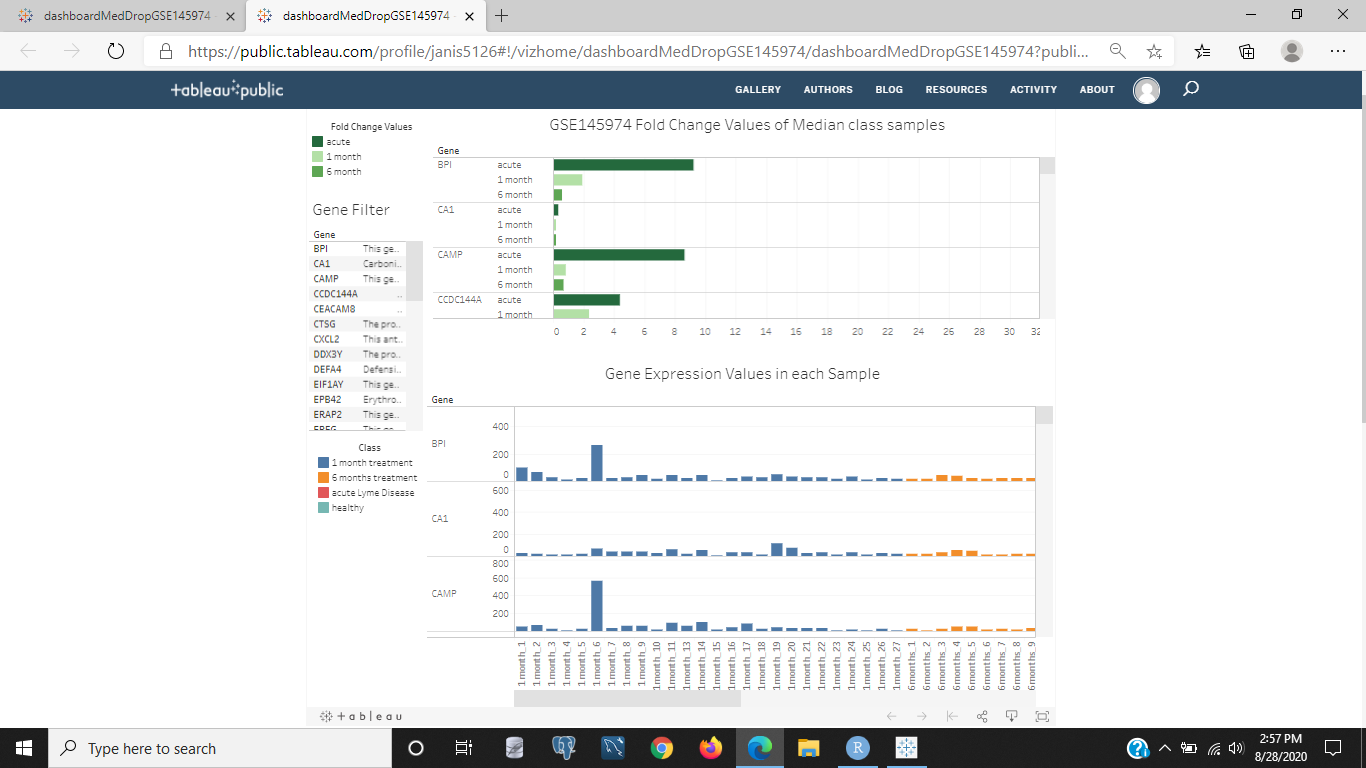


Figure 10: Dashboard of median fold change values with six samples removed. The individual samples are shown in each class as well as each top or bottom gene in fold change expression of diseased or treated to healthy ratio of medians per class values. There is a gene filter to select each gene interested in. To select more than one gene at a time use ctrl + click on each gene, to remove the selection and return all genes select the selected genes again. Even when removing the six samples that had some different gene behaviors compared to the neighboring samples within each class when using the mean values for fold change values, there are other samples in the median selected sets with odd behaviors as well that could throw off the classification algorithms. This dashboard needs to be looked through to get a list of those samples with odd behaviors and also the genes' fold change values behavior. Such as, monotastically increasing or decreasing from acute ->1 month -> 6 months, or start high, drops low, then shoots really high in gene expression after 6 months of treatment, and other behaviors. They could allow us to zoom in on target genes for lyme disease pathogenesis in these separate classes.

When it came to the machine learning to predict the accurate classes on a split of our data into 70% training samples and 30% testing samples in our de-standardized data the overall accuracy was from 20% to 54% accuracy. But when removing the most deviated samples, or samples with gene expression values the furthest away from the average of all samples within each class, the accuracy improved up to 100% accuracy when testing only one sample in each class that was the closest to the average. I also tested the data in its original format of log2 normalized values and found that the genes in this set scored up to 50% accuracy as is in overall predictions of the four classes, and removing the most deviated sample from each class and testing with a 70-30 split ratio of training samples to testing samples, the score was improved to 60%, but it had 100% accuracy on the first run predicting if the class was taking six months of antibiotics for lyme disease or not. Then, the three most deviated samples in each class was removed, except for the 6 months of treatment class that already scored 100% accuracy in precision, recall, and overall accuracy. The 6 month class remained with only one of the most deviated samples removed. The score of 30% was lower with the same 70-30 training and testing data split than it was for one or no samples that are most deviated removed. but when splitting into a 95-5 training and testing split so that the model had more samples to train on the three most deviated samples removed from three classes and one most deviated sample removed from the 6 month class, and also having at least one sample from each of the four classes in the testing set, the accuracy overall in prediction jumped to 75%, but the healthy and acute classes had 100% accuracy in precision and recall, making this model up to 100% accurate in predicting if a sample is healthy or has acute stage lyme disease without antibiotic treatment. And the 6 month class was lowered in accuracy due to it misclassifying a 1 month class as 6 month of treatment but correctly identifying the 6 month class in the testing set.

Both sets of genes that are completely different are good sets of genes to predict whether a sample is healthy or has lime disease, but the de-standardized data set with only one deviated sample removed from every class scored the best in prediction accuracy, precision, and recall. Precision is the number of correctly predicted positives as a ratio to the number of positives predicted (not the number of total positives). Recall is the number of correctly predicted positives as a ratio to the number of total true positives in the data. Recall is also called sensitiviy or true positive rate.

There was also a side machine learning experiment done on taking the median of each class samples and using that for the fold change values instead of the mean, for the ratio of acute, 1 month of antibiotics, or six months of antibiotics to the median of the healthy samples. Our classes were shortened to 'acute','1 month,','6 months,', and 'healthy' in our machine learning data tables and for the Tableau visualizations of charts and dashboards. This side experiment is referenced in the same experimental machine learning analysis on Lyme disease data, or you can also find it as a separate document on rpubs. The rpubs document for this entire analysis with a link to the side analysis is available at: <https://www.rpubs.com/janisharris/lymeDiseaeGSE145974>, and the side analysis using median data can be found at: <https://www.rpubs.com/janisharris/lymeDiseaeGSE145974part2median6samplesDropped>. The median values data had many samples skewed and none were dropped, but the highest accuracy was 42% which was less than the mean values of data as fold change values to find our top genes, and this was also a different gene set than our other two gene sets.

Thank you for viewing this blog and I appreciate your time and think you will find some use for any of the data in this document.