Tyler Iams
5/6/2018

EDA Report
Predicting Cancer Based on Mutation Profiles

Introduction

The data for this project has thus far been supplied by The Cancer Genome Atlas[1]. I have collected mutation data for four preliminary types of primary cancer cites: lung, brain, leukemia (bone marrow), and kidney. I have collected data for 400 cases total thus far, 100 cases for each type of cancer. The data collection process has once again been slower than would be preferable, but a reliable collection system has now been established. For each of the 400 cases I have three predictors:

i. gene – the name of the gene that has been mutated.
ii. genomic _dna_change – the specific base-pair in the genome that was mutated.
iii. chromosome – the chromosome that the mutation has occur upon.

Exploration

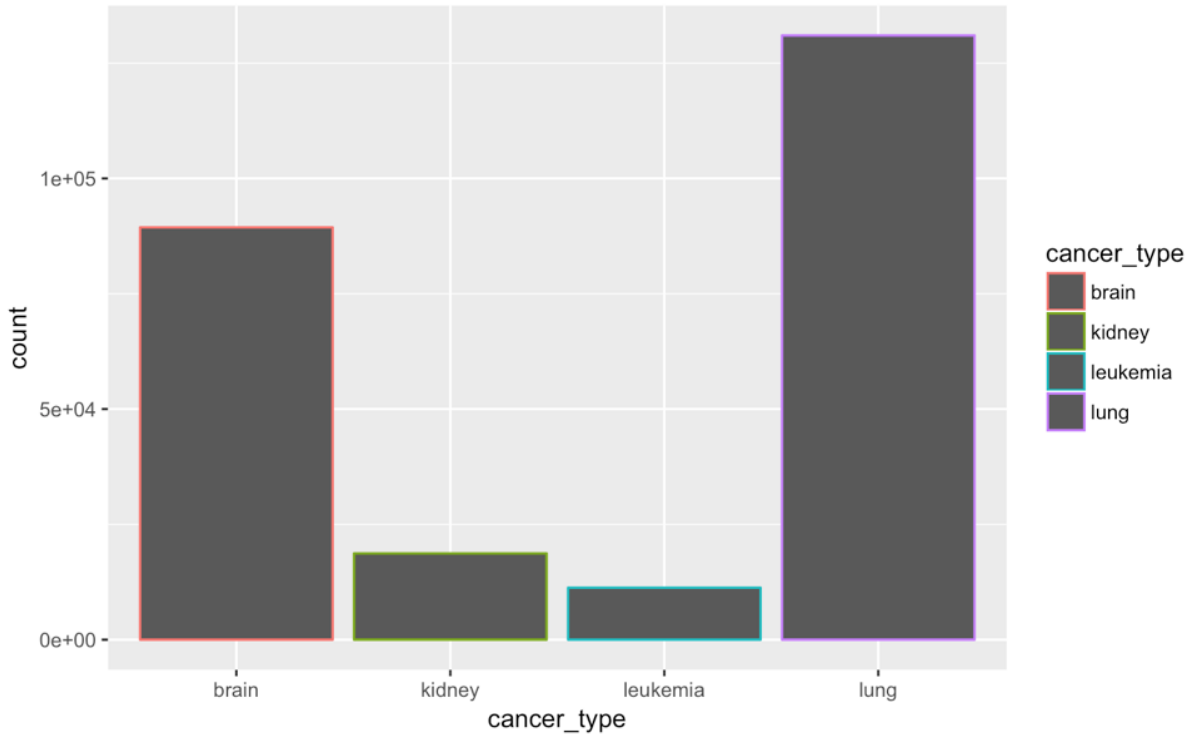
The first thing I did was explore the prevalence of mutations within each cancer type. I had an idea of the layout but felt it necessary to summarize it so it could be used to standardize the data later on:

Table 1

primary_site	num_mutations
brain	89397
kidney	18718
leukemia	11277
lung	131013

Table 1: Lung appears to have the most mutations, while leukemia has the least.

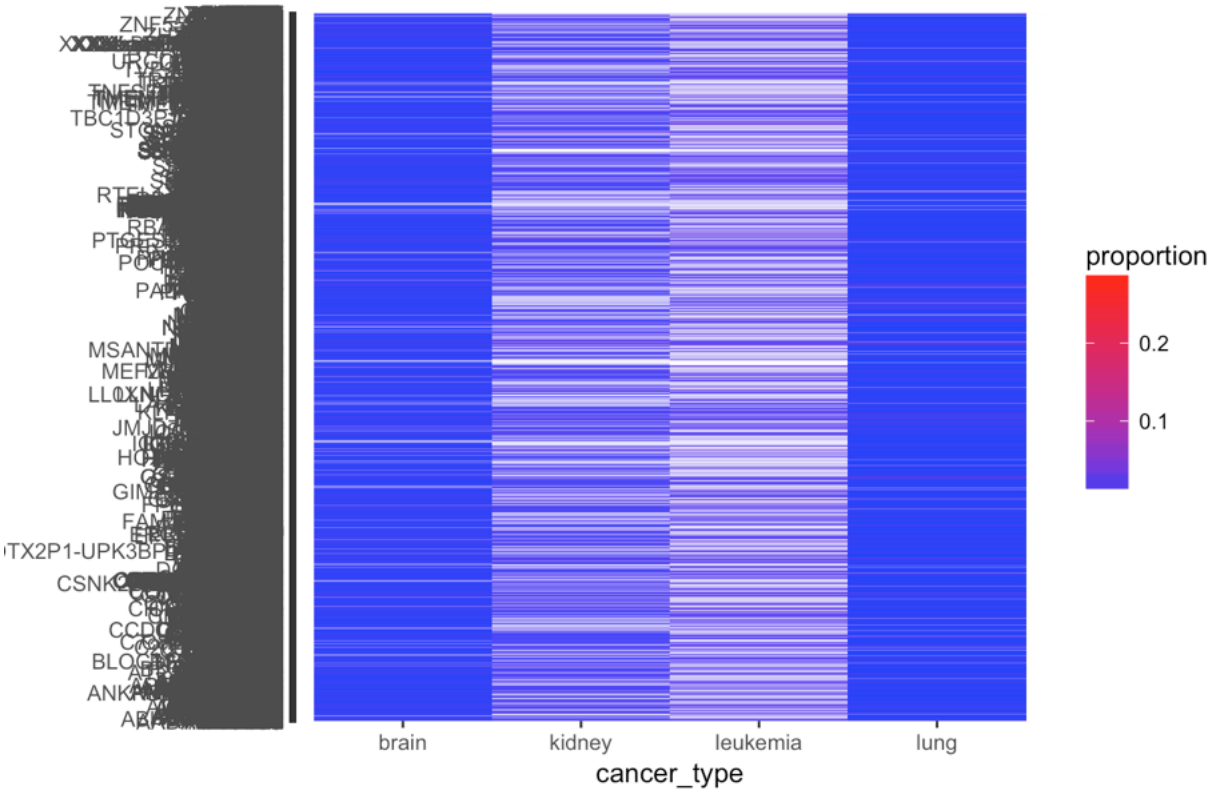
Plot 1



Plot 1: Another view of the data layout, lung being the most prevalent

Next, it was important to understand which mutations were the most prevalent while the data was grouped by cancer type. To do this, I first filtered out all of the "genomic_dna_change" data, as this caused thousands of duplicates in the "gene" field. Next, I created a field called "proportion" by calculation the prevalence of each mutation within each cancer type. This way, I was able to see that "kidney" and "leukemia" seemed to have mutations at a higher proportion than "brain" and "lung." This is shown by Plots 2 through 5:

Plot 2

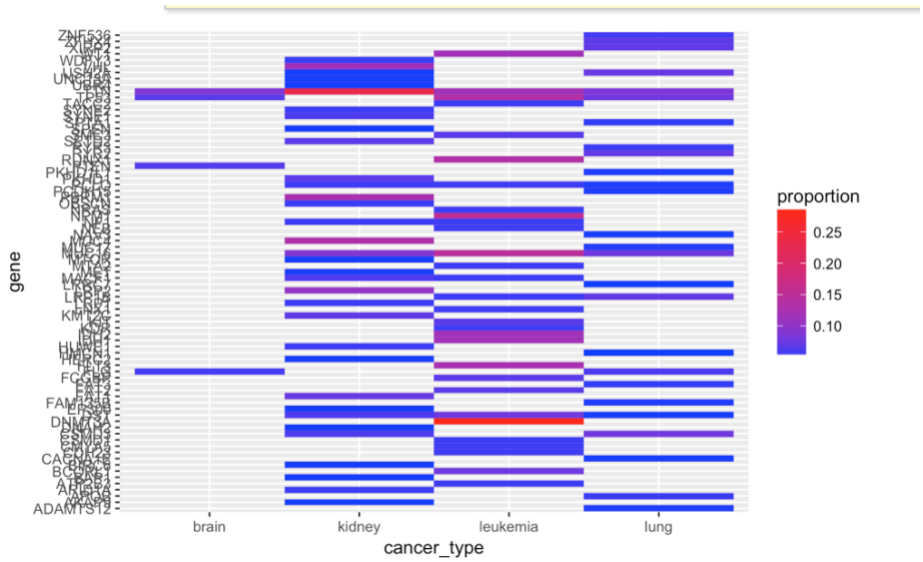


Plot 2: The y-axis is a list of all the different mutations, and unreadable.

Plot 2, though messy, can tell us a surprising amount of information about the dataset. The color blue indicates mutations that occur at a low rate, and the depth of the fill explains that there are a lot of mutations that occur at a low rate within patients with lung and brain cancer, while in patients with kidney cancer and leukemia it can be seen that there are fewer mutations that occur a lot within that population. This could be a key observation while modeling, as mutations for leukemia and kidney cancer may be more telling than those of lung and brain.

Zooming in a little bit on this find, I created Plot 3:

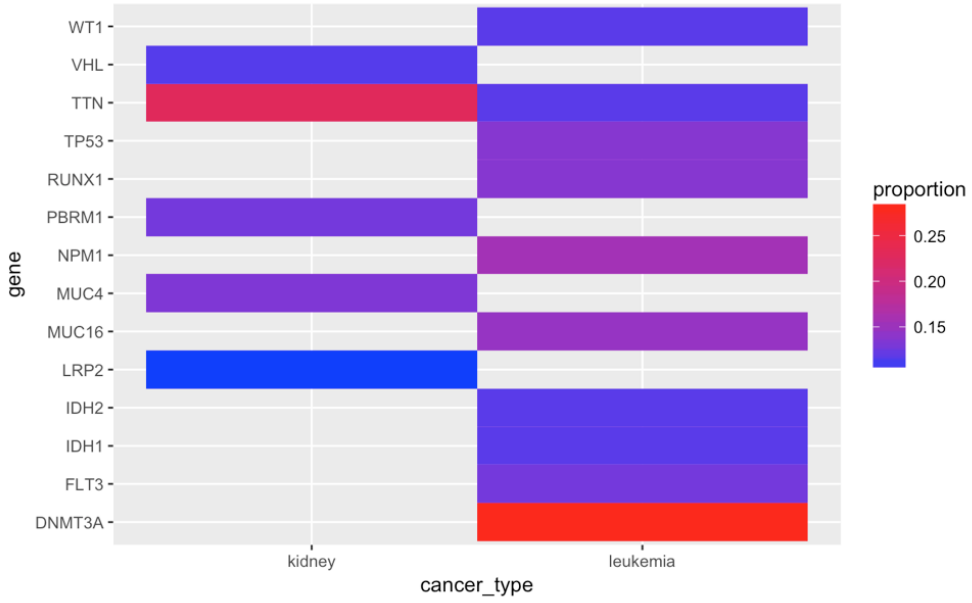
Plot 3



Plot 3: The names of the mutated genes are illegible, however the plot is telling once again.

In Plot 3 it can again be seen that mutations that occur at a high rate within patients with leukemia and kidney are more common than mutations that occur at a high rate within lung, and they are even less common in brain. One final zoom is shown by Plot 4:

Plot 4



Plot 4 shows the mutations that occurred at the highest proportions (over 10%) in patients of the various cancer types. It can be seen that patients with leukemia have the most in common in terms of their mutations

The final step in my exploration was the final step necessary to create models, and that was to create mutation profiles. This was a much larger undertaking than expected, and required a lot more computing power than I had expected. However, I now have (saved to disk), a mutation profile for each patient that I can begin modeling with. I believe I will be able to predict certain types of cancer, possibly “leukemia” vs “not leukemia,” with some accuracy.

Citation

1. The National Cancer Institute, (2018). The Genomic Data Commons Repository. https://portal.gdc.cancer.gov/repository.