Jessica Miron

Homework 6

BME 598: Applied Programming

Executive Summary

Cancer is characterized by uncontrolled cell proliferation leading to cancerous tumors. Treating cancer requires slowing or stopping this cell proliferation and killing cancer cells. The dataset provided contains information for cell proliferation across three cell lines (HA, T98G, and U251) for different perturbations of miRNA mimics and drugs. This data is being analyzed to determine viable therapeutic responses which are anti-proliferative. A combination miRNA mimic and drug which decreases glioblastoma line proliferation while keeping human astrocyte proliferation the same is the goal of this analysis. Steps for the analysis taken can be seen below.

1. Load data from the CSV file and Python packages
2. Calculate the median background value for each cell line for the three no cell conditions
3. Correct the proliferation values by subtracting the median background value for each cell line and no cell condition from each cell count value
4. Prepare the proliferation values for plotting by removing the no cell rows and changing from a wide to long format. Replicates are changed to just contain the cell line name.
5. Plot proliferation using violin plots. Each plot contains each perturbation for the cell line as a violin plot with the four replicates as the data values. Three plots are created with 39 perturbations each.
6. Calculate the fold change for each perturbation for each cell line compared to the negative control. Divide the median proliferation perturbation value by the negative control proliferation value
7. Calculate the T-statistic and p-value for each perturbation for each cell line compared to the negative control. Each group has four replicates which are compared to the negative control's four replicates. 35 perturbations are present in the final dataset after removing the negative controls.
8. Put all fold change, T-statistic, and p-values into a Pandas Data Frame by perturbation and cell line before writing to a CSV file for further comparison.
9. Using the violin plots and the outputs of the statistical tests in the CSV, complete a biological interpretation on which perturbations and cell lines were statistically different and what each difference means.

Biological Interpretation

To compare treatments, violin plots were created of each perturbation for each cell line to visually see differences. Additionally, the fold change, T-statistic, and p-value comparing each perturbation per cell line to its negative control were found to compare numerically. To determine if treatments increased or decreased proliferation, the violin plots and fold change can be used. For the violin plots, each treatment is compared to its negative control to observe if the

treatment caused the median to be greater or lower than the negative control median. For the HA cell line (see Figure 1), there are four groups to observe: NC_Mimic, NC_Mimic.2, NC_siRNA, and V.DMSO negative controls. These groups correspond to miRNA mimics, a second group of miRNA mimics, siRNA, and drug treatments. Starting with the NC_Mimic group miR_23a_Mimic has the greatest increase in median compared to the negative control. This means miR_23a_Mimic increased cell proliferation the most in both the group and over all treatments. miR_29a_Mimic has the highest peak value due to its larger data spread. miR_223_Mimic has the lowest median compared to the negative control in the group with miR_15b_Mimic and miR_133b_Mimic both having the lowest value due to data spread. For the NC_Mimic.2 group, miR_331_3p_Mimic has the highest median and miR_486_3p_Mimic has the lowest median. For the NC_siRNA group, IKZF1_siRNA has the highest median, though it is very close to the control median, and ETV6_siRNA has the lowest median. For the V.DMSO group, which are all drug treatments, Imatinib_Drug has the highest median and Romidepsin_Drug has the lowest median though there are three other drug treatments with similarly low medians. After observing these treatments in the violin plots, Table 1 can be used to numerically compare the treatments. Each cell line in Table 1 has three values which are color coded. For the HA cell line there are 17 statistically significant differences, two of which increased proliferation. The only significant increases were for treatments miR_23a_Mimic and miR_29a_Mimic which were observed to be the greatest medians in the cell line. The greatest significant decreases were all drug treatments, with seven out of nine drug treatments showing a significant decrease. The numeric values in Table 1 match with the observations made about Figure 1.

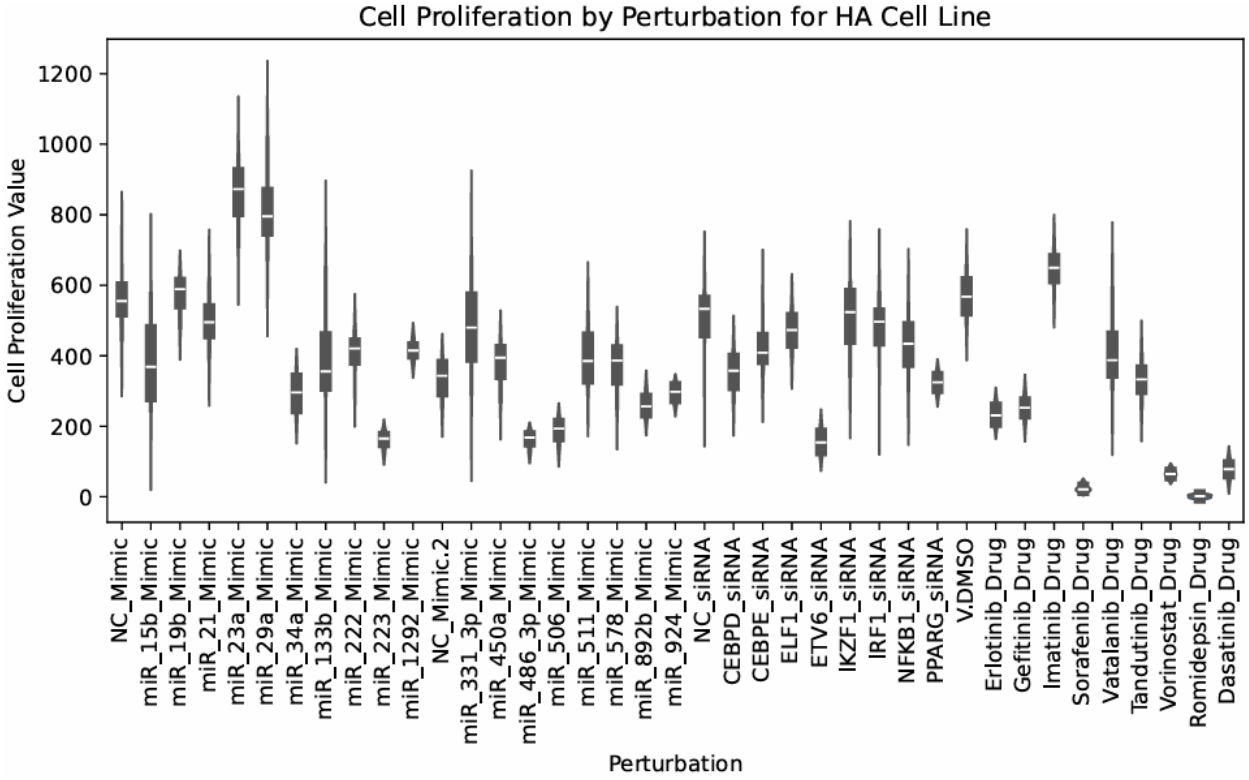

Figure 1: Cell proliferation by perturbation for HA cell line. The y-axis is the cell proliferation value while the x-axis shows which treatment is being shown. NC_Mimic, NC_Mimic.2, NC_siRNA, and V.DMSO are the four negative controls. Treatments to the left of each control before the next control are compared to the left control for data analysis.

For the T98G cell line, Figure 2 shows the violin plots of the different perturbations. For the NC_Mimic group, miR_23a_Mimic has the greatest median and miR_34a_Mimic has the smallest median. For the NC_Mimic.2 group, no treatments have a median greater than the control and miR_892b_Mimic has the smallest median. For the NC_siRNA group, IRF1_siRNA has the greatest median of all treatments and ETV6_siRNA has the smallest median of the group. For the V.DMSO group, no treatment has a median greater than the control and Romidepsin_Drug has the smallest median of all treatments though three other drug treatments have similar medians. Comparing to Table 1, the T98G cell line has 22 statistically different treatments. Of the 22 statistically different treatments, six are significant increases. Four are miRNA mimics and two are siRNA treatments. The greatest increase is the IRF1_siRNA treatment which had the greatest median of all treatments. Looking at significant decreases, all nine drug treatments had a significant decrease in proliferation compared with the negative control with Romidepsin_Drug having the largest decrease. All drug treatments are also classified as anti-proliferative with Romidepsin_Drug also being the most anti-proliferative.

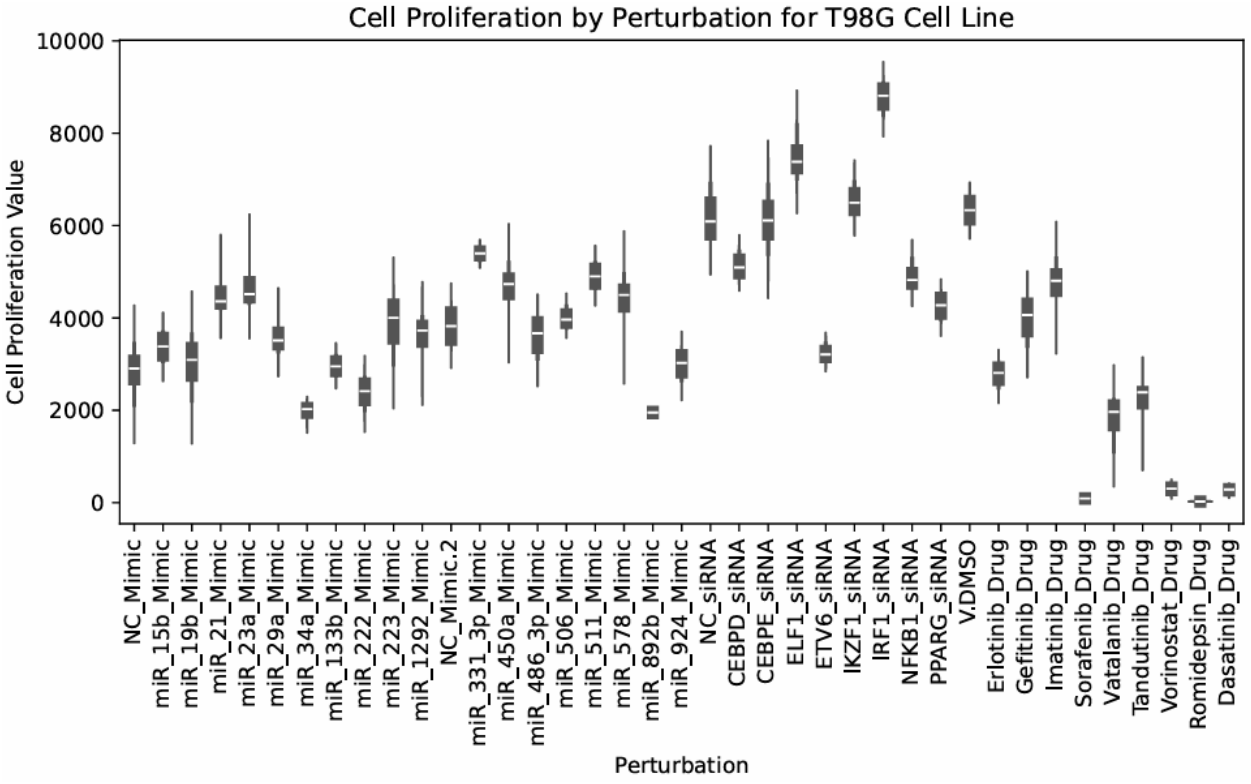

Figure 2: Cell proliferation by perturbation for T98G cell line. The y-axis is the cell proliferation value while the x-axis shows which treatment is being shown. NC_Mimic, NC_Mimic.2, NC_siRNA, and V.DMSO are the four negative controls. Treatments to the left of each control before the next control are compared to the left control for data analysis.

For the U251 cell line, Figure 3 shows the violin plots of the different perturbations. For the NC_Mimic group, miR_19b_Mimic, miR_23a_Mimic, and miR_29a_Mimic have greatest median with all three medians being very similar. miR_34a_Mimic has the smallest median of the group. For the MC_Mimic.2 group, miR_331_3p_Mimic has the greatest median and miR_486_3p_Mimic has the smallest median. For the NC_siRNA group, ELF1_siRNA and IRF1_siRNA have the greatest medians with both being very similar. ETV6_siRNA has the smallest median of the group. For the V.DMSO group, Imatinib_Drug has the greatest median of all treatments and Romidepsin_Drug has the smallest median of all treatments. Comparing to Table 1, the U251 cell line has 26 statistically significant differences. Eight are significant increases with five being miRNA mimics, two being siRNA treatments, and one being a drug treatment. The greatest increase in proliferation is for miR_331_3p_Mimic. The greatest decrease in proliferation is for Romidepsin_Drug. Seven of the nine treatments are classified as anti-proliferative with Imantinib_Drug increasing proliferation and Gefitinib_Drug significantly decreasing from the control but not having a small enough fold change to be anti-proliferative.

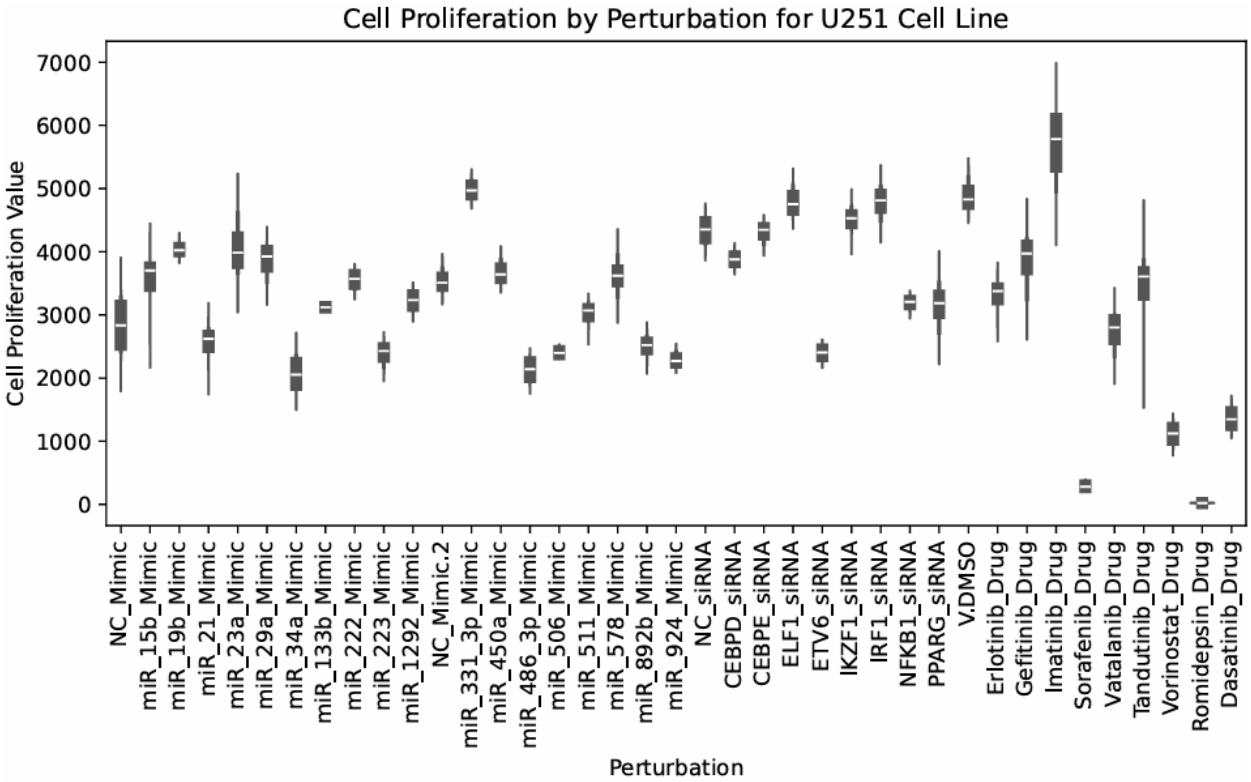

Figure 3: Cell proliferation by perturbation for U251 cell line. The y-axis is the cell proliferation value while the x-axis shows which treatment is being shown. NC_Mimic, NC_Mimic.2, NC_siRNA, and V.DMSO are the four negative controls. Treatments to the left of each control before the next control are compared to the left control for data analysis.

Table 1: A table of all perturbations used across the three cell lines. Each cell line has the fold change (*.FC), T-statistic (*.stat), and p-value (*.pvalue) listed. Each column is color coded to make viewing easier. The fold change is colored green under 0.8 (named as anti-proliferative from the paper) and red over 1 (meaning proliferation increased). The T-statistic is colored the greenest for the most negative value (greatest decrease in proliferation

compared to the negative control) and the reddest for the most positive value (greatest increase in proliferation compared to the negative control). The p-value is colored yellow if the value is below 0.05 (named as significant from the paper).

	HA.FC	HA.stat	HA.pvalue	T98G.FC	T98G.stat	T98G.pvalue	U251.FC	U251.stat	U251.pvalue
miR_15b_Mimic	0.66396396	-1.9207303	0.10316161	1.16397109	1.74080461	0.13236305	1.30697264	2.18559009	0.07150812
miR_19b_Mimic	1.06126126	0.02053075	0.98428571	1.06366139	0.40483245	0.69963838	1.42065313	5.68801054	0.00127403
miR_21_Mimic	0.89189189	-0.9081241	0.398818	1.50086029	4.84712985	0.00285998	0.92586055	-1.2018096	0.2747154
miR_23a_Mimic	1.57207207	3.81939842	0.00876686	1.55333792	4.97803963	0.00250663	1.40688438	4.2028906	0.00566762
miR_29a_Mimic	1.43333333	2.86177638	0.02873476	1.20836201	2.32844967	0.05876528	1.3858782	4.24579207	0.00540514
miR_34a_Mimic	0.53333333	-4.615672	0.00363179	0.69735031	-3.0999198	0.02111864	0.72480141	-3.1847798	0.01896096
miR_133b_Mimic	0.64054054	-1.5724112	0.16691576	1.01514109	0.39941613	0.70341589	1.10150044	1.36204684	0.22209862
miR_222_Mimic	0.75765766	-2.5417942	0.04397438	0.83103923	-1.4588843	0.19487909	1.26107679	3.31767941	0.01605062
miR_223_Mimic	0.2972973	-7.459816	0.00029926	1.37783895	2.39866085	0.05339221	0.85719329	-2.0866749	0.0819744
miR_1292_Mimic	0.74684685	-2.7354991	0.03393759	1.28286304	2.01681558	0.09029787	1.14245366	1.76374639	0.12822883
miR_331_3p_Mimic	1.40145985	1.78147466	0.12512097	1.41218779	8.13871713	0.00018493	1.41609687	15.3612653	4.81E-06
miR_450a_Mimic	1.15182482	0.90729209	0.39922394	1.23839414	2.43350148	0.05091945	1.03703704	1.40360948	0.21001637
miR_486_3p_Mimic	0.49051095	-5.5402459	0.00145923	0.95880738	-0.8845369	0.41044827	0.60982906	-13.664794	9.54E-06
miR_506_Mimic	0.5649635	-4.3694358	0.00472192	1.03726952	0.87012716	0.41767788	0.68319088	-14.686618	6.26E-06
miR_511_Mimic	1.12408759	1.32439293	0.23358366	1.28180986	4.86984756	0.00279481	0.87421652	-4.9955568	0.00246322
miR_578_Mimic	1.12846715	0.67462337	0.52504236	1.17614751	1.49741273	0.1849315	1.03105413	0.51911568	0.62226419
miR_892b_Mimic	0.74744526	-2.0500442	0.08623675	0.51065777	-10.116879	5.42E-05	0.71851852	-9.9139027	6.08E-05
miR_924_Mimic	0.87007299	-1.1740344	0.28486666	0.79102916	-3.5945856	0.01144066	0.6465812	-14.607418	6.46E-06
CEBPD_siRNA	0.6713615	-2.097352	0.08077288	0.83622034	-3.6463316	0.01075329	0.89128936	-4.5414568	0.00392718
CEBPE_siRNA	0.76713615	-0.7748006	0.46788264	1.00353009	-0.2116804	0.8393636	0.99896575	-0.2687613	0.79711575
ELF1_siRNA	0.88826291	-0.2855153	0.7848497	1.21221575	3.50866289	0.01269201	1.09216272	3.67496654	0.01039267
ETV6_siRNA	0.28920188	-5.4857295	0.00153517	0.52680404	-10.635417	4.07E-05	0.55263158	-19.910055	1.04E-06
IKZF1_siRNA	0.98403756	0.112435	0.91414584	1.06641491	1.08195717	0.32082713	1.04125488	1.33125337	0.23145218
IRF1_siRNA	0.93239437	-0.2645814	0.80018605	1.44717182	8.31689645	0.00016391	1.10618249	3.18277444	0.01900908
NFKB1_siRNA	0.81502347	-0.7773668	0.46647721	0.79172482	-4.3647803	0.00474581	0.73718685	-11.856514	2.18E-05
PPARG_siRNA	0.60938967	-2.7859023	0.0317487	0.70183072	-6.6655361	0.00055164	0.73201563	-6.3509611	0.00071432
Erlotinib_Drug	0.40740741	-8.9632316	0.00010773	0.44372534	-21.593365	6.44E-07	0.69939921	-10.407096	4.61E-05
Gefitinib_Drug	0.44532628	-8.1870609	0.00017893	0.64119968	-9.2574428	8.98E-05	0.82266418	-4.5223532	0.00400754
Imatinib_Drug	1.14550265	1.64919827	0.15020153	0.75848461	-5.5266741	0.00147773	1.19846696	2.58254564	0.04162677
Sorafenib_Drug	0.03703704	-15.5881	4.41E-06	0.01428571	-50.518428	4.04E-09	0.05821421	-45.761667	7.30E-09
Vatalanib_Drug	0.68342152	-2.1469683	0.07542125	0.31026046	-16.300331	3.39E-06	0.58027761	-12.438646	1.65E-05
Tandutinib_Drug	0.58730159	-5.1108625	0.00219784	0.37687451	-15.408595	4.72E-06	0.74756578	-4.6113748	0.00364819
Vorinostat_Drug	0.11375661	-14.377605	7.09E-06	0.04751381	-46.743163	6.42E-09	0.23254609	-31.987237	6.21E-08
Romidepsin_Drug	0.0026455	-16.350265	3.33E-06	0.0038674	-51.06208	3.79E-09	0.0045577	-49.189934	4.73E-09
Dasatinib_Drug	0.13844797	-13.376314	1.08E-05	0.04483031	-47.828184	5.60E-09	0.27905531	-29.537614	9.98E-08

Comparing the three cell lines there are multiple trends between non-cancerous and cancerous lines as well as key differences between cancerous lines. While the non-cancerous line had the fewest number of treatments with significant differences, it shows more anti-proliferative treatments that are miRNA mimics than the cancerous lines. The first group of miRNA mimics show the greatest difference of anti-proliferative treatments. The cancerous lines for the same treatments had more pro-proliferative treatments. For the siRNA treatments, the non-cancerous line has two significant differences compared to six and six for the cancerous lines. However, the non-cancerous line had all anti-proliferative or fold changes under 1 compared to the cancerous lines having multiple treatments that increased proliferation. For drug treatments, the non-cancerous line had two non-significant differences while all treatments for both cancerous lines were significant. The drug treatments generally followed the same trend for every line with

Romidepsin_Drug being the most anti-proliferative followed by Sorafenib_Drug, Vorinostat_Drug, and Dasatinib_Drug. These trends show that cancerous lines become more proliferative under miRNA and siRNA treatments compared to the non-cancerous line becoming less proliferative. The most obvious difference between cancerous lines is Imatinib_Drug. For the T98G line the drug was anti-proliferative but for HA and U251 it was pro-proliferative. This signals that some part of the Imatinib_Drug is able to stop the T98G line but increases the U251 line showing how important it is to know the cell line before cancer treatment starts.

The combinations of Vatalinib and miR-892b and Romidepsin and miR-506 should be prioritized. Combinations were chosen because they decrease cell proliferation for T98G and U251, cancerous lines, while minimizing cell proliferation changes for HA, non-cancerous line. miR-892b shows a significant decrease in cell proliferation for T98G and U251 but not a significant change for HA. Vatalinib also has a significant decrease for T98G and U251 but not for HA. This combination would decrease cell proliferation for the tumor cells inhibiting growth while not changing cell proliferation for the healthy brain cells. Romidepsin was the drug shown to have the largest decrease in cell proliferation for all cell lines, though the cancerous cell lines had a much greater negative T-statistic. miR-506 significantly decreased HA and U251 cell lines and had a non-significant increase in T98G. If treatment requires a more aggressive solution, this combination would greatly decrease cell proliferation for U251 and T98G to fight the cancer more aggressively. This would come at the cost of a decrease in healthy cell proliferation which could be harmful but necessary depending on cancer severity and aggression.