MicroArray Based Tumor Classification - Programmer and Biologist Roles

Luke Zhang

Introduction:

This project focused on using modern bioinformatics techniques in order to perform in depth analysis on microarray data. The two roles for this project which were performed were the Programmer and Biologist, both which served to contextualize the microarray results.

The Programmer role focused more on preprocessing the data and quality control, using the Robust Microarray Averaging algorithm to normalize all the microarray data collected for the study as well as producing various quality control metrics. This serves to ensure that downstream data analysis produces results which are both accurate and interpretable, and ensures that any significance gleaned from further analysis would not be attributed to bias or mishandling of the data.

The Biologist role focused on contextualizing the results of the study as well as providing a framework to investigate the significance of the results. The metrics computed in this step help to identify gene expression patterns which may be interesting enough to investigate further, more specifically looking at significant differences in tumor expression data for each subtype.

Methods - Programmer:

The data processing and quality control step was started by reading CEL files containing the microarray samples using the Affy (v1.70) R package, obtainable through Bioconda Manager (v3.14). After microarray samples were read and stored in R dataframes, the package AffyPLM (v1.72) was used in order to produce relative log expression (RLE) and normalized unscaled standard error scores (NUSE) of the microarray samples, plotting the medians of each sample in a histogram. Following the creation of RLE and NUSE median histograms, the ComBat module from the SVA package (v3.42) was used to correct for batch effects in conjunction with relevant metadata. Finally, PCA was performed on scaled/centered corrected data, and the first and second principle components were plotted against one another.

Methods - Biologist:

The biologist step began by matching ProbeIDs with Gene Symbols for a precomputed differential expression matrix by using the hgu133plus2db package, removing duplicate matches by selecting the ProbeID with the lower adjusted p-value. The differential expression matrix used contained 22903 samples. The top thousand most up and down-regulated genes were then selected, with the top ten of each stored in a separate table for further analysis. The two thousand genes mentioned above were then used in conjunction with KEGG, GO, and Hallmark genesets

for the calculation of hypergeometric statistics and p-values comparing overlap with each gene set. The statistics were calculated using a Fisher's t-test, with p-values for multiple hypotheses adjusted using the Benjamini-Hochberg (FDR) procedure. The top three gene sets sorted by nominal p-value for each set of hypothesis testing were then reported.

Results - Programmer:

The resulting data pre-processing and quality control yielded a corrected, RMA-normalized list of genes which numbered at 54675. The quality of the data looked to be reasonably good, with the distributions of median RLE (Fig 1) and NUSE (Fig 2) values centered around 0 and 1 respectively. The distribution of median RLE values were unimodal and relatively symmetrical while the distribution of median NUSE values were unimodal with a noticeable skew towards the right. The range of median NUSE values is also notably smaller than the range of median RLE values, with the NUSE values ranging from 0.98 to 1.06 while median RLE values ranged from -0.15 to 0.2.

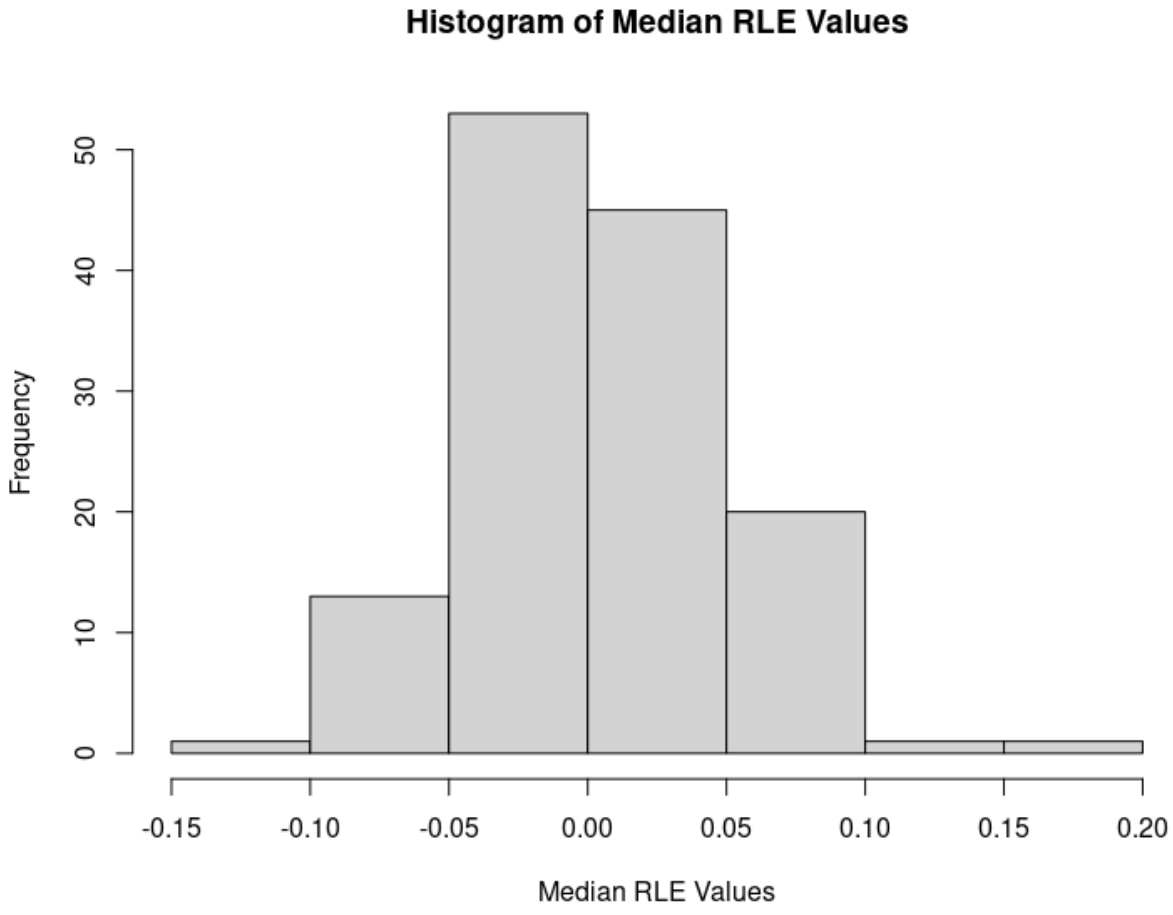

Figure 1: Distribution of Median RLE Values. The values were centered around 0 with a relatively symmetric distribution.

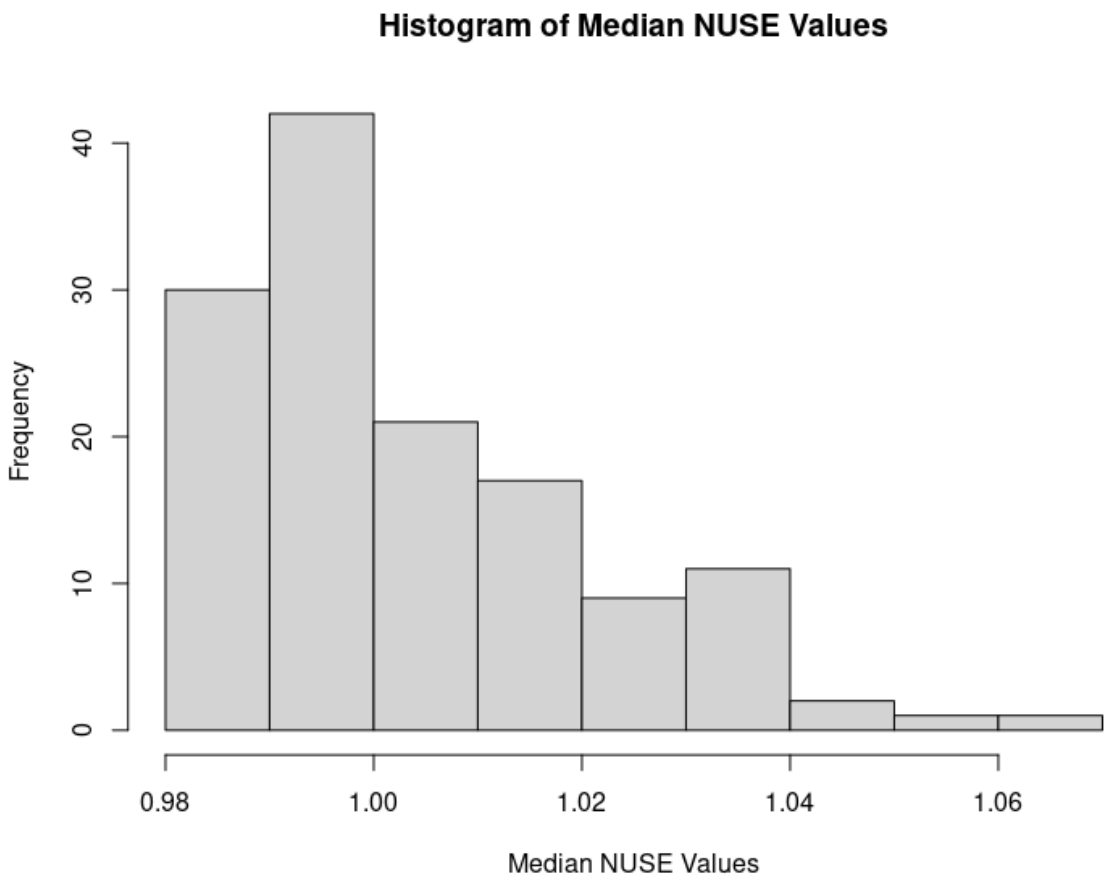

Figure 2: Distribution of Median NUSE Values. The distribution of NUSE values skews noticeably to the right, with the center around 1/1.01.

The results produced by principal component analysis also looked to be acceptable, with an even spread centered around zero (Fig 3). It is important to note that outliers are present in Figure 3, which manifest themselves as points which deviate significantly from the center of the plot. The percent variation calculations for PC1 and PC2 were 11.4% and 8.4% respectively, which is larger than expected as these two principal components account for around 20% of total variance out of 134 principal components. This could be an indication that the data was not processed properly.

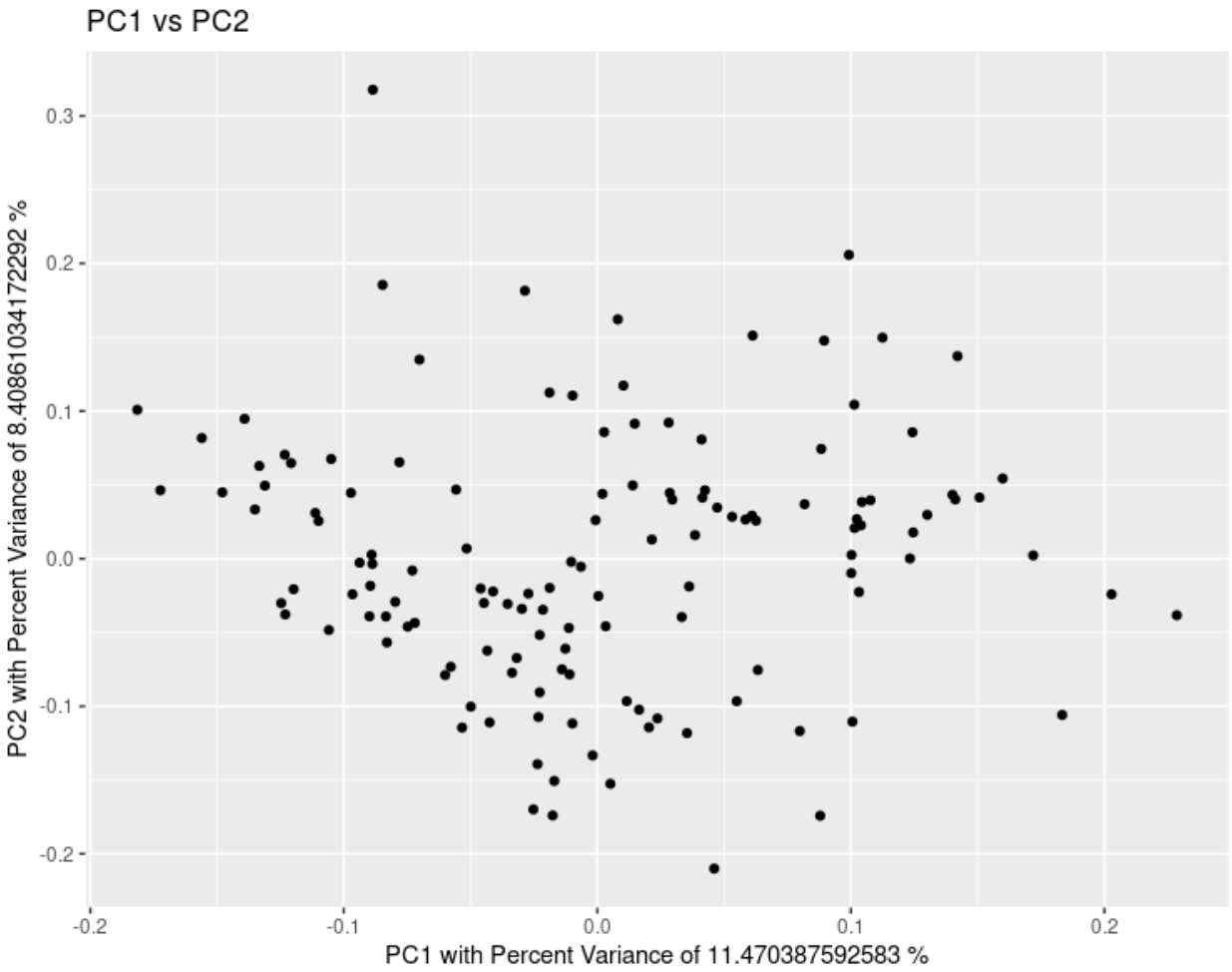

Figure 3: Plot of PC1 vs PC2, with percent variance for each component listed on the x-axis. High percent variance could indicate that the data was not correctly normalized.

Results - Biologist:

The results of this part of the analysis take the form of lists of up-regulated and down-regulated gene sets. Using a pre-computed differential expression matrix, gene symbols were attached to correct probeset IDs and filtered to produce the top ten most up and down-regulated probesets (Table 1a/b). Although these lists were based purely on the pre-computed t-statistic, and not the native or adjusted p-value, all probe sets in both lists had p-values which fell below the threshold to be considered statistically significant. One factor which may have resulted in differing results from the paper could be how multiple probeset ID and gene symbol matches were filtered, as for this project duplicates with the lowest adjusted p-values were kept. The original paper may have had access to datasets with the correct gene symbols or had another way of matching the probeset IDs to the gene symbols. This would become more relevant for the next step of the analysis.

	PROBE_ID	SYMBOL	T-Statistic	P-Val	P-Adj
1	202291_s_at	MGP	20.90408	2.059140e-43	5.895060e-40
2	202363_at	SPOCK1	20.97745	3.862302e-43	8.845829e-40
3	204457_s_at	GAS1	22.16718	6.426347e-45	2.943653e-41
4	207266_x_at	RBMS1	22.65447	2.565982e-47	2.938434e-43
5	213413_at	STON1	21.03553	3.832251e-40	4.388502e-37
6	219778_at	ZFPM2	20.59668	1.424420e-35	3.106999e-33
7	223122_s_at	SFRP2	23.30672	1.345746e-48	3.082162e-44
8	225242_s_at	CCDC80	21.27925	2.009370e-43	5.895060e-40
9	226930_at	FNDC1	20.95565	2.547358e-43	6.482461e-40
10	227059_at	GPC6	20.88544	1.142413e-42	2.378608e-39

Table 1a. Top 10 up-regulated probe sets. The t-statistic was already calculated in the differential expression matrix, and used to determine how much the probe set was regulated, with higher t-statistics representing up-regulation and vice versa.

	PROBET_ID	SYMBOL	T-Statistic	P-Val	P-Adj
1	203240_at	FCGBP	-13.78812	2.686508e-25	1.087086e-23
2	205489_at	CRYM	-12.80386	1.148261e-24	4.311248e-23
3	211715_s_at	BDH1	-13.41787	2.633650e-25	1.067584e-23
4	214106_s_at	GMDS	-12.80637	1.136680e-21	2.995785e-20
5	218189_s_at	NANS	-12.68730	5.388846e-24	1.878550e-22
6	220622_at	LRRC31	-13.54314	1.535462e-26	7.176876e-25
7	222764_at	ASRGL1	-12.60942	1.616677e-23	5.366198e-22
8	227725_at	ST6GALNAC1	-13.13729	1.154818e-22	3.403963e-21
9	234008_s_at	CES3	-12.58871	2.431507e-24	8.825485e-23
10	235350_at	C4orf19	-12.60836	1.747309e-22	5.033941e-21

Table 1b. Top 10 down-regulated probe sets. The t-statistic was already calculated in the differential expression matrix, and used to determine how much the probe set was regulated, with higher t-statistics representing up-regulation and vice versa.

The next set of lists produced the top three most enriched gene sets for each geneset type, comparing the GO, HallMark, and KEGG genesets. Each of the genesets referred to previously is a publicly available, annotated set of gene symbols. GO, or GeneOntology.org, seeks to develop a computational model for gene function and provides genesets with evidence-based annotations for pathways and functions. KEGG, referring to the Kyoto Encyclopedia of Genes and Genomes, contains annotated pathways for high-level biological functions. HallMark datasets, originating from MSigDB, provide the user with a "refined" gene set which emphasizes genes which show specific biological state/function and displays cohesive expression. Each gene set type was used in conjunction with the aforementioned up and down-regulated genes in a Fisher's T-Test to find the most enriched genesets. HallMark originally contained 50 genesets, GO originally contained 10402 genesets, and KEGG originally contained 186 genesets. After adjusting for a P-value < 0.05, the number of enriched genesets found for HallMark, GO, and KEGG respectively were 74, 217,10019. The top three most enriched gene sets, found by filtering for the highest t-stat estimates, can be seen in Tables 2-4 below.

The resulting enriched gene sets differed greatly from those found in the original paper, as there were no matches between the enriched genes reported by the paper and those found in this project. This is cause for concern as this may mean that the hypergeometric testing was done incorrectly, the incorrect gene sets were used, or there may have been a bug during filtering. This could further be affirmed by the absurdly large Fisher T-Stat Estimates which were reported.

GO GeneSet Name	P-Value	Adj. P-Value	Fisher T-Stat Estimate
GOBP_ANTIBODY_DEPENDENT_CELLULAR_CYTOTOXICITY	3.12885111566792e-14	1.22585e-12	14502.81
GOBP_PEPTIDYL_LYSINE_OXIDATION	3.12885111566792e-14	1.22585e-12	14502.81
GOBP_RESPONSE_TO_CORTISOL	3.12885111566792e-14	1.22585e-12	14502.81

Table 2: Top 3 enriched gene sets for GO genes.

HallMark GeneSet Name	P-Value	Adj. P-Value	Fisher T-Stat Estimate

HALLMARK_NOTCH_SIGNALING	9.960144057749972e-09	5.242181e-08	94.04707
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	5.906584655195594e-153	5.906585e-151	157.55195
HALLMARK_ANGIOGENESIS	1.501599033350245e-28	2.502665e-27	408.46877

Table 3: Top 3 enriched gene set for HallMark genes.

KEGG GeneSet Name	P-Value	Adj. P-Value	Fisher T-Stat Estimate
KEGG_TAURINE_AND_HYPOTAURINE_METABOLISM	1.813044768360048e-05	1.204380e-04	479.1602
KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_CHONDROITIN_SULFATE	2.217138801110091e-19	9.164174e-18	910.5392
KEGG_ECM_RECEPTOR_INTERACTION	1.389018248275553e-48	5.167148e-46	170.7591

Table 4: Top 3 enriched gene set for KEGG genes.

Discussion

The programmer role focused on data preprocessing and quality control, which yielded RLE and NUSE median value distributions as well as a PCA plot comparing the first two principal components. The median value distributions did not present any significant outliers or anomalies, while the PCA plot displayed a spread of points which aligned with the centering and scaling which were done to the data.

The biologist role focused on finding enriched gene sets, using three gene set types (GO, KEGG, and Hallmark) through the use of hypergeometric statistical tests. The resulting enriched gene sets did not match those found in the paper and may be indicative of flaws in the analysis steps, or the code which would lead to abnormally large test statistic values presented.

References:

1. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Fléjou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P, Boige V. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5):e1001453. doi: 10.1371/journal.pmed.1001453. Epub 2013 May 21. PMID: 23700391; PMCID: PMC3660251.

2. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. doi: 10.1073/pnas.0506580102. Epub 2005 Sep 30. PMID: 16199517; PMCID: PMC1239896.