Re-analysis of a subset of Colorectal Tumor Microarray-based Classification

Kyrah Kotary (Analyst), Marina Natividad (Programmer), Vishwa Talati (Biologist), Brad Fortunato (Data Curator)

Introduction:

Colorectal cancer (CRC) is among the most prevalent cancers among both sexes and responsible for over 600,000 every year. Pathology testing determines severity, treatment and prognosis, but confers no prediction on chance of recurrence. Although CRC biomarkers have been identified, we are yet to find a molecular signature that confers an advantage over pathological testing for a more accurate patient stratification and remission surveillance. Marisa et al. (2013) gathered clinical and pathological data of 750 CRC patient samples in various stages. The quality of 566 allowed for subsequent gene expression profiling (GEP) analysis. Samples were then split between discovery and validation with an additional 905 samples from public databases to the validation set. Through GEP, the authors were able to further subclassify the samples into 6 molecular subtypes (C1-4) with distinct molecular signatures and found that C4 and C6 subtypes were more likely to relapse than the others. In the present work, we reproduce the results of the GEP analysis of subtypes C3 and C4 with a total of 134 samples.

Data:

Fresh frozen tumor samples were obtained from the The French national Cartes d'Identite´ des Tumeurs (CIT) program[2], with a total cohort of 750 patients featuring stage I to stage IV colon cancer (CC). Of the total 750 samples, 566 were deemed to meet quality standards to perform Gene Expression Profiling(GEP) upon. Said samples were further divided into an n = 443 discovery set and n = 123 validation set, with an additional 906 samples obtained from other public databases (GSE13067, GSE14333[4], GSE13294[5], GSE17536/17537[6], GSE18088[7], GSE26682[8], and GSE33113[9]). Said datasets fulfill quality checks including; GEP data attained through use of an analogous platform (Affymetrix U133 Plus2.0 chips) along with CEL files with DNA mutation and patient prognosis data. GEP was performed on the 556 CC approved samples on a Affymetrix U133 Plus 2.0 chips platform. For 19 of the samples, adjacent non cancerous tissue was also tested. Afterwards the datasets were normalized through the use of R (Packages affy and SVA). Of the 750 CC samples, 464 were determined to be high enough quality to be analyzed through array-based comparative genomic hybridization.

As for our project, first we created a central project folder on the SCC for us to consolidate our analyses. We located the missing sample GSM971958, downloaded the .CEL.gz file, then moved said file to our shared project folder. From there, a symbolic link was created to allow ease of access to the other samples files located in the samples directory.

Sample files were also uploaded to our git repository for further ease of access.

Methods:

Statistical Methods

Raw Affimetrix data was downloaded in the form of CEL files. To reduce unwanted variation caused by factors not related to the biology of interest, the raw-expression data were normalized with the Robust Multi-Array Average (RMA) method using the Affy package in Bioconductor, giving the same empirical distribution of intensities to each array. Next, we calculated the relative precision of the expression across arrays; the relative log expression (RLE) was plotted to evaluate whether the normalization succeeded in removing unwanted variation. To determine whether there was a relative difference in quality between arrays, the affyPLM package was used to calculate the normalized unscaled standard error (NUSE). The data was corrected for batch effects using the sva package in Bioconductor and a principal component analysis was performed to examine for outliers.

Gene Set Enrichment and Biological interpretation :

We mapped the probeset ID to gene symbols using hgu133plus2.db package of Bioconductor using the select function. For the symbols that map to multiple probe id, we selected the ones with most significant p value (i.e. minimum p_adjusted) and removed the remaining. Top 1000 up and down regulated genes were chosen based on highest and lowest t-values. We downloaded the KEGG, GO and hallmark gene set collection from MsigDB with .gmt files having gene symbols for gene set enrichment. These were loaded using the GSEABase package of Bioconductor. The top 1000 gene sets were enriched using Fisher test and adjusted the values using Benjamini Hochberg method.

Results

Quality Control

Proof of a successful normalization are presented in Figures 1-4. After plotting for PC1 vs PC2 (Fig 5) ,separation between the two subtypes is hinted as dots samples aggregate up and down an imaginary diagonal.

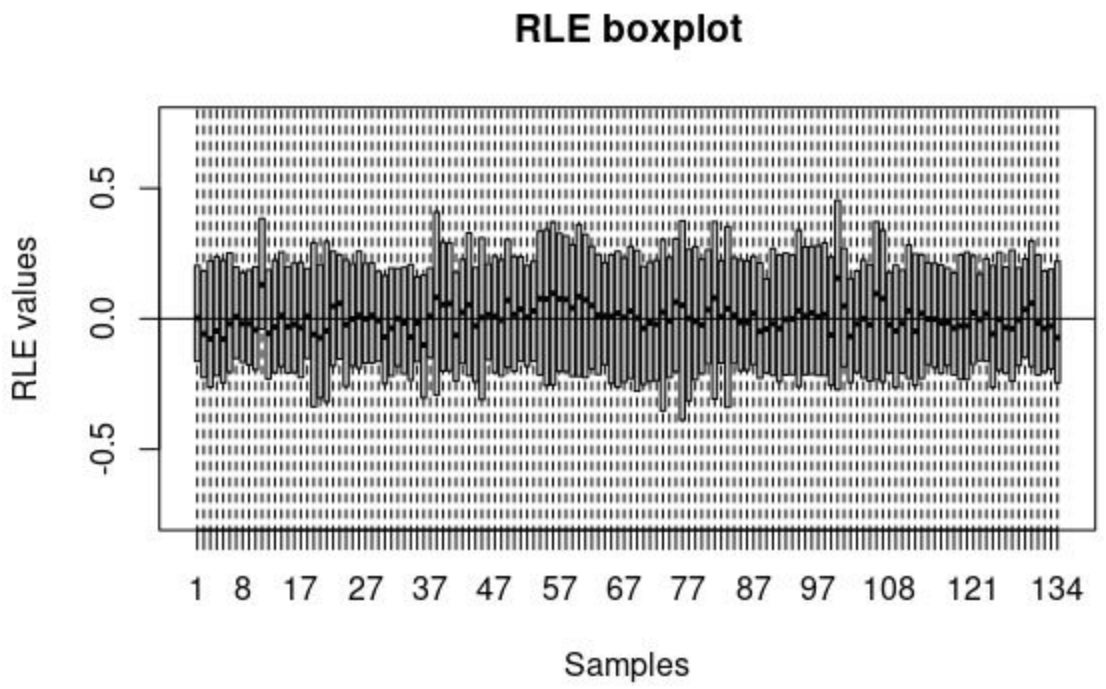

Figure 1. RLE plot for C3 and C4 samples. Lower quality arrays have a median not centered in 0.

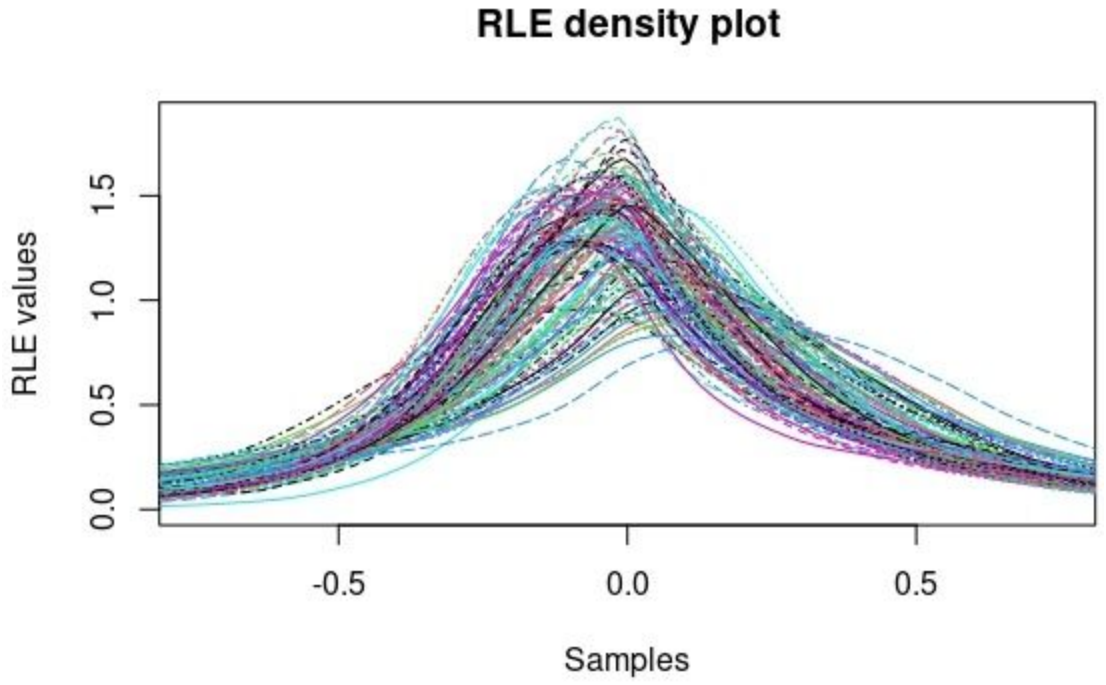

Figure 2. RLE density plot for C3 and C4 samples. By being centered close to 0, the plot provides proof that unwanted variation was successfully removed.

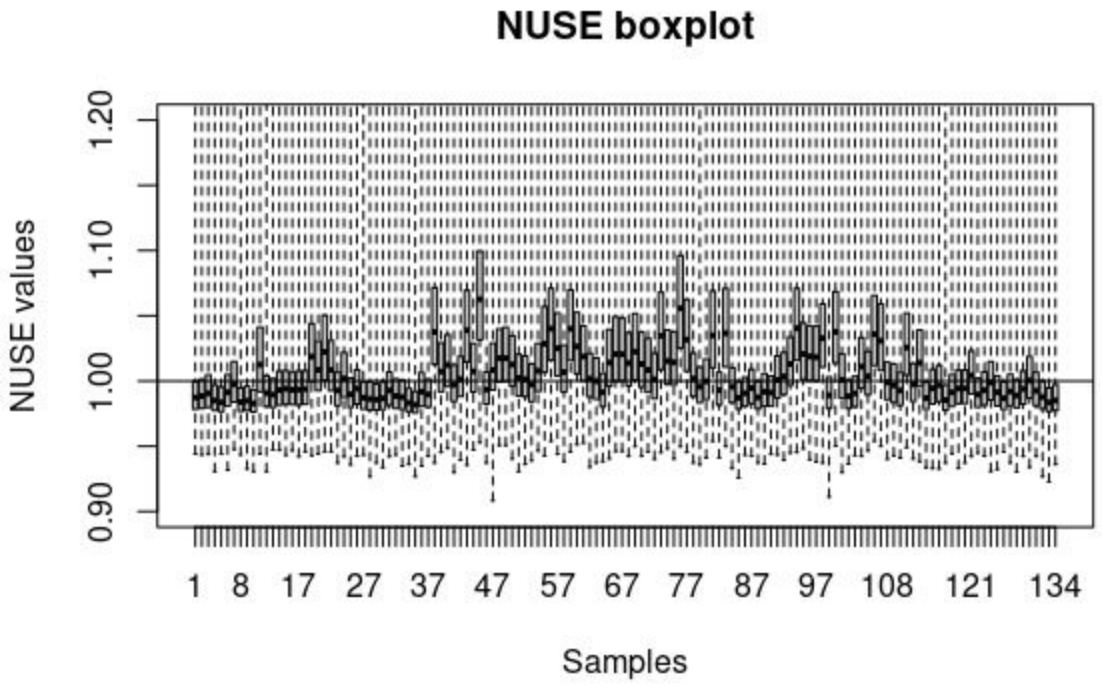

Figure 3. NUSE boxplot for C3 and C4 samples. In NUSE, the standard errors are normalized to have a median of 1, therefore arrays of lesser quality differ from others in the dataset by being elevated or more spread out relative to other arrays.

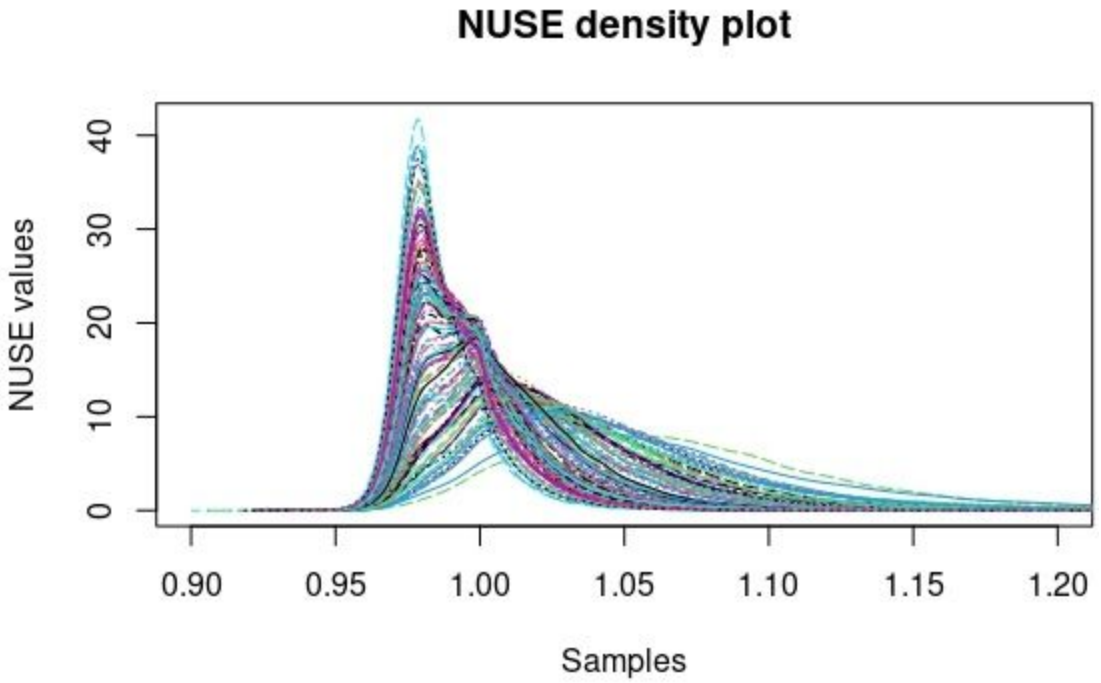

Figure 4. NUSE plot for C3 and C4 samples.

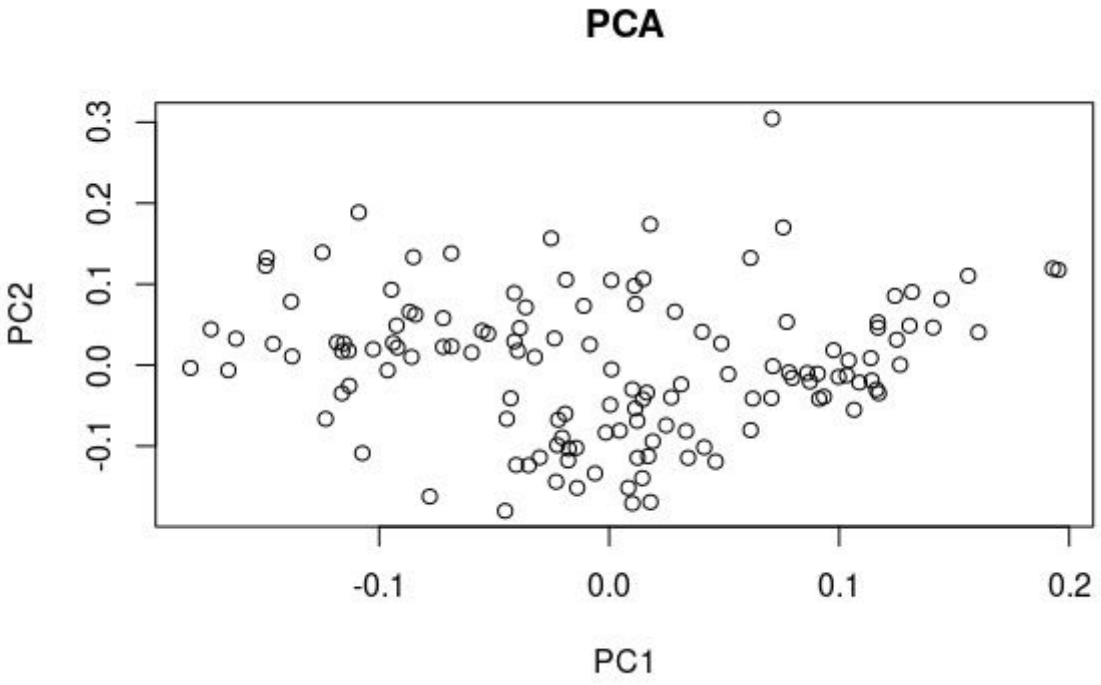

Figure 5. Plot of PC1 vs PC2. Subclassification starts becoming evident.

Noise Filtering and Dimensionality Reduction

After normalization and adjustment for batch effects, the expression matrix contained 54675 gene probes. Noise filtering and dimensionality reduction were completed to select genes for clustering analysis. Filtering was based on three metrics defined by Marisa *et al.* (2013), which included expression in at least 20% of the samples, having a variance significantly different from the median variance of all probe sets (defined as having $p < 0.01$), and having a coefficient of variation greater than 0.186. A total of 15518 genes passed all three thresholds.

Hierarchical Clustering and Subtype Discovery

After filtering the genes, those that passed the appropriate thresholds were clustered using a hierarchical clustering method, in order to find novel relationships within the data. The data was divided into two clusters, with cluster 1 having 78 samples and cluster 2 having 56 samples. This clustering method was employed to group the samples based on the similarity of their gene expression patterns, which is useful for identification of biological signatures of interest. The heatmap in figure 1 shows the clusters and the distribution of the C3 and C4 subtypes within them. According to this heatmap, neither cluster displays a clearly dominant subtype.

A Welch t-test was used to identify differentially expressed genes between the two clusters. This analysis revealed that a total of 10392 genes were differentially expressed at an adjusted p value less than 0.05 between the clusters.

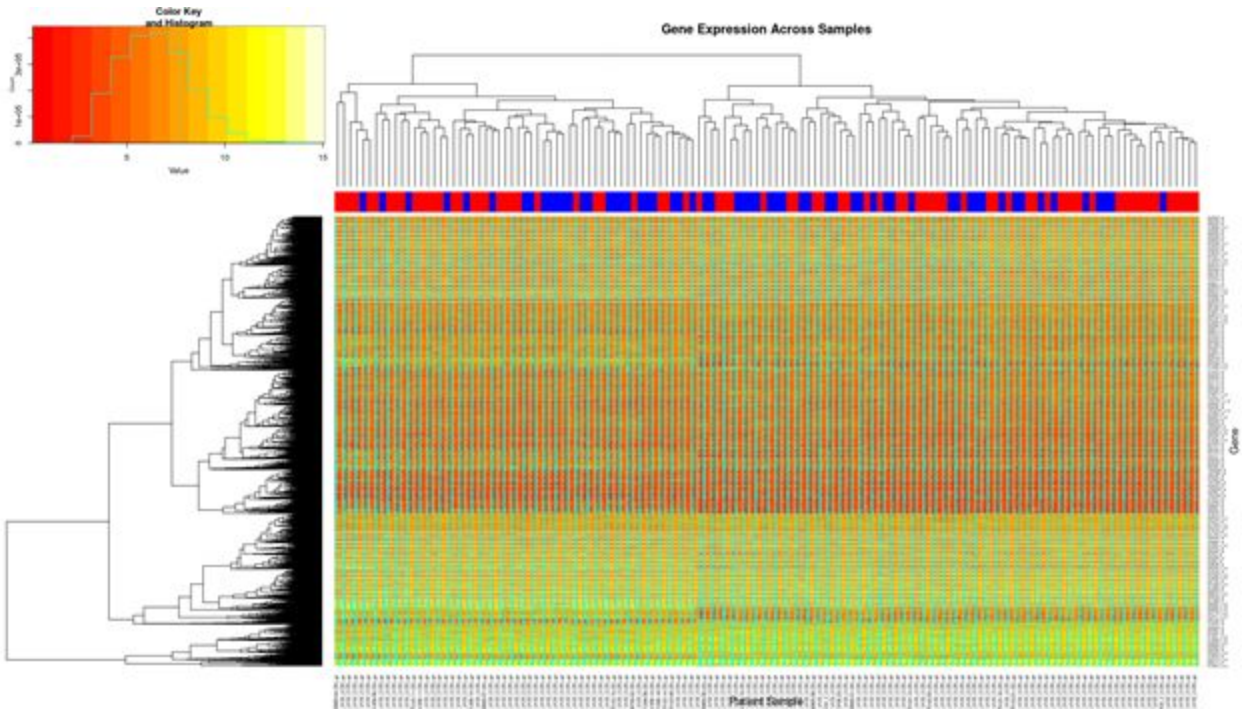

Figure 6. Heatmap showing the expression of each gene across the 134 patient samples. Red indicates that the sample belongs to the C3 subtype and the C4 subtype is indicated with blue.

Biological Interpretation and Analysis:

From MsigDB gene set collection, we used three gene sets namely: Hallmark, Gene Ontology(GO) and Kyoto Encyclopedia of Genes and Genomics(KEGG) to enrich the top 1000 up and down regulated genes as shown in tables. Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. It has a collection of 50 gene sets.All gene sets derived from Gene Ontology. It has a collection of 10271 gene sets. KEGG has a collection of 186 Canonical Pathways gene sets. After performing gene set enrichment analysis, we get 762 enriched gene sets with adjusted p value < 0.05.

PROBEID	t	p	p_adjust	SYMBOL
1568598_at	-0.8004	0.426857	0.426857	KAZALD1
64432_at	-0.819	0.416253	0.416307	MAPKAPK5-AS1
214779_s_at	-0.821	0.415121	0.415228	SGSM3
203201_at	-0.82671	0.411904	0.412089	PMM2
227686_at	-0.82935	0.410419	0.410638	OXNAD1
205313_at	-0.83078	0.409619	0.409909	HNF1B
217942_at	-0.83204	0.408913	0.40923	MRPS35
221627_at	-0.83629	0.40654	0.406934	TRIM10
218272_at	-0.83634	0.40651	0.40693	TTC38
228361_at	-0.83641	0.406475	0.406921	E2F2

Table 1: Top 10 upregulated genes based on t-statistic values

PROBEID	t	p	p_adjust	SYMBOL
218660_at	-1.25129	0.216021	0.391003	DYSF
219431_at	-1.25304	0.215392	0.391003	ARHGAP10
227313_at	-1.25805	0.213586	0.391003	CNPY4
232217_at	-1.25931	0.21314	0.391003	CALHM5
201885_s_at	-1.26255	0.211967	0.391003	CYB5R3
44702_at	-1.26348	0.211644	0.391003	SYDE1
218231_at	-1.26401	0.211449	0.391003	NAGK
209721_s_at	-1.26614	0.2107	0.391003	IFFO1
45749_at	-1.26751	0.210208	0.391003	RIPOR1
219183_s_at	-1.27045	0.209168	0.391003	CYTH4

Table 2: Top 10 down regulated genes based on t-statistic values

The top 3 gene sets that are significantly enriched in the GO collection are shown in Table 3. These gene sets include circulatory system development which help fight diseases and help stabilize body temperature and pH to maintain homeostasis, mitochondrion that is the site of tissue respiration and mitochondrial matrix that contains the enzymes of the tricarboxylic acid cycle.

For hallmark gene set collection, the top 3 gene sets are shown in Table 4. These gene sets include genes that encode components of apical junction complexes and also the proteins that are involved in metabolism of fatty acids oxidative phosphorylation.

Table 5 has enriched gene sets from KEGG gene set collection. These gene sets are related to Valine, leucine and isoleucine degradation, propanoate metabolism and regulation of actin cytoskeleton.

Geneset name	**p value**	**estimate**	**exp**	**BH**
GO_MITOCHONDRION	2.03E-31	3.468211	UP	4.18E-27
GO_MITOCHONDRIAL_MATRIX	3.64E-27	5.952035	UP	3.74E-23
GO_CIRCULATORY_SYSTEM_DEVELOPMENT	2.60E-18	2.756925	Down	1.78E-14

Table 3: Enriched Gene Sets from GO collection based on p-value

Geneset name	**p value**	**estimate**	**exp**	**BH**
HALLMARK_APICAL_JUNCTION	9.49E-08	3.579542	Down	9.49E-06
HALLMARK_FATTY_ACID_METABOLISM	2.96E-07	3.909969	UP	1.48E-05
HALLMARK_OXIDATIVE_PHOSPHORYLATION	6.83E-07	4.162199	UP	2.15E-05

Table 4: Enriched Gene Sets from Hallmark collection based on p-value

Geneset name	p value	estimate	exp	BH
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	7.37E-08	8.423876	UP	2.17E-05
KEGG_PROPANOATE_METABOLISM	1.16E-07	12.60599	UP	2.17E-05
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	4.42E-07	3.462724	Down	5.48E-05

Table 5: Enriched Gene Sets from KEGG collection based on p-value

Discussion

We consider it worth reminding the reader the two assumptions made before a microarray normalization are that the expression of most genes will not change due to the biological process being studied and that the number of upregulated genes is almost equal to the number of downregulated genes. Although an RLE can tell whether or not unwanted variation prevails, it is unable to detect whether biological variation was removed from the data.

Differing results can be partially attributed to several factors lending to our methodology. Instead of performing analysis on all of the 6 subtypes, focus was placed on only 2 subtypes (C3 & C4). By reducing the sample cohort, we most likely altered the non-biological variance of our sample and although the RMA method is robust, our data post-normalization is bound to be different from that of Marisa *et al.* (2013).

Biological interpretation is given by the Gene set enrichment process. The findings from the paper demonstrate that secreted frizzled-related protein 2 (SFRP2) and growth arrest-specific 1 (GAS1) included in top deregulated genes are markers of the aggressiveness of CC cells and may constitute potential therapeutic targets. However, that was not the case with our findings because the t-values generated were not in accordance with the t-value generated in the research paper. Amongst the top 10 up regulated genes from our result, MAPKAPK5-AS1 is associated with hepatocellular carcinoma and amongst top 10 down regulated genes, SYDE1 and ARHGAP10 which are related to GPCR signaling(GTPase activator) that plays important role in cancer progression.

Conclusions:

Based on our findings, gene set enrichment of top 1000 genes show pathways related to protein binding, lipid binding and metabolism of cytochrome P450 are down regulated and pathways related to translation and differentiation were up regulated.

References

1) Carlson, M., & Obenchain, V. (2020). *Creating select Interfaces for custom Annotation resources*. Retrieved from website: https://www.bioconductor.org/packages/release/bioc/vignettes/AnnotationForge/inst/doc/MakingNewAnnotationPackages.pdf
2) CIT Program | Carte d'Identité des Tumeurs - Accueil. *Cit.ligue-cancer.net*. Retrieved from https://cit.ligue-cancer.net/
3) select function | R Documentation. (2021). Retrieved February 23, 2021, from Rdocumentation.org website: https://www.rdocumentation.org/packages/ensembldb/versions/1.4.7/topics/select
4) GEO Accession viewer. Ncbi.nlm.nih.gov. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75317
5) GEO Accession viewer. Ncbi.nlm.nih.gov. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13294
6) GEO Accession viewer. Ncbi.nlm.nih.gov. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29623
7) GEO Accession viewer. Ncbi.nlm.nih.gov. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18088

8) GEO Accession viewer. Ncbi.nlm.nih.gov. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26682

9) GEO Accession viewer. Ncbi.nlm.nih.gov. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33113

10) GSEA | MSigDB. (2020). Retrieved February 23, 2021, from Gsea-msigdb.org website: http://www.gsea-msigdb.org/gsea/msigdb/index.jsp

11) GSEABase. (2021). Retrieved February 23, 2021, from Bioconductor website: https://bioconductor.org/packages/release/bioc/html/GSEABase.html

12) Gene Ontology Consortium. (2020). AmiGO 2: Term Details for “circulatory system development” (GO:0072359). Retrieved February 24, 2021, from Geneontology.org website: http://amigo.geneontology.org/amigo/term/GO:0072359

13) Database, G. (2021). GeneCards - Human Genes | Gene Database | Gene Search. Retrieved February 24, 2021, from Genecards.org website: https://www.genecards.org/