Gene Expression Classification of Colon Cancer Defines Six Molecular Subtypes with Distinct Clinical, Molecular and Survival Characteristics [Expression]

Teresa Rice, Arushi Shrivastava , Maha Naim

INTRODUCTION

With the help of clinical advancements within the realms of screening and treatment of colorectal cancer, scientists have gained greater understanding of the etiology and pathogenesis of the disease. Despite this, however, unanswered questions remain within the context of relapse of colorectal cancer (CRC). Pathological staging allows for examination of the extent of the disease and indicates potential associated prognoses and treatment options. Unfortunately, this system does not accurately predict recurrence among CRC patients undergoing surgery. To address this problem, prior studies have explored the predictive value of gene expression profiles with regards to the prognosis of CRC. However, to date, there are no existing conclusions which offer clinical relevance to improve the current disease stratification [1-4]. One study, *Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value*, took the approach of using gene expression profile analyses to build a standard molecular classification of CRC [5]. In addition, associations between identified molecular subtypes, clinical and pathological factors, common DNA alterations, and prognosis were evaluated [5]. The bioinformatic techniques used in the study allowed for identification of six unique molecular subtypes which were analyzed via multivariate analysis and according to biological relevance. Comparative analysis of such genomic data permits greater room for interpretation of molecular interactions and identification of distinct patterns.

DATA

The data being used for this study was obtained from the CIT, where they collected the 750 tumor samples out of which 566 samples pass the quality filter and this set was used to model. Also here they consider the *Homo sapiens* as an organism and the Experiment type was Expression profiling by array. The 443 samples were considered as the discovery set and 123 samples were considered as the validation set. Along with that they also utilized the publicly available data, i.e. seven datasets for validation.

Among all the collected sample from CIT, 598 samples were analyzed for mRNA expression profiles using Affymetrix U133plus2 chip and, among these, 463 could also be analyzed for DNA alteration profiles using the CGH Array (CIT-CGHarray V6). The 585 tumors were divided into a discovery dataset of 443 CC and a validation dataset of 123 CC and 19 non-tumoral dataset.

The source of the data is the GEO database. Where I had searched and downloaded the CEL.files. The GEO ID is GSE39582. As it was mentioned in the paper. Under this ID they have all the data available, also consider the CEL.file from other GEO Id as they have also used publicly available dataset for validation.

Here is the link to the repository (“ https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582 “). Moreover, being the data curator, some of the CEL files were provided and my role was to find the missing files. Then upload the files to the SCC server and make it available for the other members of the group.

METHODS

For data quality control, the data had to be cleaned then a couple quality control tests were done. The data was first normalized, followed by computations for the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores. A summary of the RLE and NUSE scores for the microarray samples have been produced in the form of two histograms. RLE scores with a median around 0 are considered to be good quality, and NUSE scores greater than 1 are considered to be bad quality, according to these metrics our data should be high quality. A dot plot is made of PC1 vs PC2 in order to identify outliers, however, no significant outliers are identified and therefore not removed. The R package `affy` was used in order to read the raw CEL files. The `rma()` function (also from the R package `affy`) was then used to normalize the affymetrix samples together. A histogram of median RLE (Figure 1) and NUSE (Figure 2) for each sample has been prepared (See end of paper). These histograms have been set to bin=35 in order to reveal more detail. The RLE and NUSE score are on the x-axis on their respective plots and a count is on the y-axis. The RLE scores appear to have a normal distribution centered over 0 while the NUSE scores have a right skew (positive skew) where the distribution is centered around 0.75. There was a small hiccup with uploading datasets at the beginning but once all needed files were available data QC proceeded smoothly. By using `summary()` on the results of the `prcomp()` the standard deviation, proportion of variance, and cumulative proportion are now available for PC1 through PC134. From this data, a PC1 vs PC2 plot was prepared (Figure 3), the proportion of variance was also used in the PC1 vs PC2 plot in order to calculate percent variability of each PC. Although analysis data for C3 and C4 were lacking, if we were to color the PC1 vs PC2 plot there would be visible clustering of C3 and C4 (Sup.1 was provided by a classmate and not produced by this group).

RESULTS

	Probe Set ID	Gene Symbol	T-statistic	P-value	P-adjusted
1	223122 s at	SFRP2	23.3067211763392	1.34574609683018e-48	3.08216228557017e-44
2	207266 x at	RBMS1	22.6544687338796	2.56598171130962e-47	2.93843395670621e-43
3	204457 s at	GAS1	22.1671770316659	6.42634713647706e-45	2.94365256933468e-41
4	225242 s at	CCDC80	21.2792504769876	2.00937034394141e-43	5.89506026244625e-40
5	213413 at	STON1	21.0355349737235	3.83225094273489e-40	4.38850216707286e-37
6	202363 at	SPOCK1	20.9774483602018	3.86230162174608e-43	8.84582940428504e-40
7	226930 at	FNDC1	20.9556455792225	2.54735834837182e-43	6.48246091697332e-40
8	202291 s at	MGP	20.9040763739065	2.05913994234685e-43	5.89506026244625e-40
9	227059 at	GPC6	20.8854379181402	1.14241335593203e-42	2.37860846281012e-39
10	219778 at	ZFPM2	20.5966763561598	1.42441995208585e-35	3.10699906310688e-33

Table 1. Top 10 up-regulated probe sets. List of the top 10 up-regulated probe sets as sorted by t-statistic value. Probe sets correspond to their appropriate gene symbols. Associated t-statistic, nominal p-value, and adjusted p-value columns are shown.

In order to perform further analysis of the probesets, probeset IDs were mapped to their appropriate gene symbol. Duplicates and missing values were removed to maintain high quality data. After sorting the data by descending t-statistic value, only the most variant probes were retained according to p-adjusted values. The top 10 up- and down-regulated genes were gathered. Results are shown in Tables 1 and 2. The top up- and down-regulated genes are SFRP2 and CES3, respectively.

Analysis of the Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and Hallmark genesets was performed. Descriptions and total amount of genesets in each collection are listed in Table 3. Of the total genesets, the number of significant gene sets per database are outlined in Table 4 after obtaining only the sets with a p-value less than 0.05. Of the 186 genesets in the KEGG collection, 124 were significant. GO consisted of the largest collection of genesets, where 1,165 of the 10,271 sets were significant. Hallmark had 50 total genesets, where only 27 were significant. Using the collected data, the Fisher test results were computed by taking into account the differentially expressed genes which were both present and not present in the set. Similarly, genes which were not differentially expressed but both present and not present in the set were accounted for in the function. Enriched pathways with p-values less than 0.05 were retained. Results are shown in Table 5 where the top 3 enriched pathways for each

geneset is listed. Each pathway was up-regulated. Pathways pertaining to cell communication including receptor interaction and focal adhesion were up-regulated for the KEGG geneset. Structural and motility-related pathways pertaining to the extracellular matrix were enriched for the GO geneset. Cancer signaling and muscle generation pathways including epithelial-mesenchymal transition were highly enriched for the Hallmark dataset.

	Probe Set ID	Gene Symbol	T-statistic	P-value	P-adjusted
1	234008 s at	CES3	-12.5887118741914	2.43150719509699e-24	8.82548483190277e-23
2	235350 at	C4orf19	-12.6083645510835	1.74730923207874e-22	5.0339406076257e-21
3	222764 at	ASRGL1	-12.609419567884	1.61667747603512e-23	5.36619771501918e-22
4	218189 s at	NANS	-12.6873035668861	5.38884648736051e-24	1.87855024505354e-22
5	205489 at	CRYM	-12.8038558167288	1.14826066436869e-24	4.31124819607148e-23
6	214106 s at	GMDS	-12.8063700961857	1.13667950829839e-21	2.99578489971898e-20
7	227725 at	ST6GALNAC1	-13.1372941938637	1.15481787328417e-22	3.40396315982333e-21
8	211715 s at	BDH1	-13.4178737311676	2.63365002377776e-25	1.06758383176251e-23
9	220622 at	LRRC31	-13.5431416078249	1.53546219465606e-26	7.17687564167506e-25
10	203240 at	FCGBP	-13.7881155288664	2.68650768873274e-25	1.08708631793367e-23

Table 2. Top 10 down-regulated probe sets. List of the top 10 down-regulated probe sets as sorted by t-statistic value. Probe sets correspond to their appropriate gene symbols. Associated t-statistic, nominal p-value, and adjusted p-value columns are shown.

Database	Total Number of Genesets	General Description [6]
KEGG	186	Subset of curated gene set collection, C2. Consists of canonical pathways.
Gene Ontology	10,271	Collection of ontologies which support biologically relevant annotation of genes and their products.
Hallmark	50	Pose a summary and representation of specific biological processes or phenomena. Display genes with coordinate expression from gene set overlaps.

Table 3. Geneset databases. 3 geneset databases were used to compute the Fisher's test. The total number of genesets within each collection is specified. A general summary of each database is also provided.

	KEGG	GO	Hallmark	Total Number of Sets
Number of Significantly Enriched Genesets	124	1165	27	1316

Table 4. Significantly enriched genesets. The number of significantly enriched genesets was found for each geneset database. Collections were sorted by having a BH value, or adjusted p-value, of less than 0.05. With these values, the total number of significantly enriched genesets was calculated.

	Geneset Name	Nominal p-value	Estimate	Expression Type	BH (adjusted p-value)
1	KEGG_ECM_RECEPTOR_INTERACTION	1.53E-15	9.06081994	Up	5.71E-13
2	KEGG_FOCAL_ADHESION	5.59E-14	4.47893181	Up	1.04E-11
3	KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	3.41E-11	8.853274424	Up	4.23E-09
4	GO_COLLAGEN_CONTAINING_EXTRACELLULAR_MATRIX	1.91E-56	8.530390217	Up	3.93E-52
5	GO_EXTRACELLULAR_MATRIX	8.33E-55	7.023951264	Up	8.55E-51
6	GO_EXTRACELLULAR_STRUCTURE_ORGANIZATION	1.05E-43	6.84950103	Up	7.22E-40
7	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	1.96E-70	16.44610997	Up	1.96E-68
8	HALLMARK_MYOGENESIS	1.67E-13	4.603308269	Up	8.35E-12
9	HALLMARK_COAGULATION	2.33E-12	5.491127765	Up	7.76E-11

Table 5. Top 3 enriched pathways for each geneset. The top 3 biological pathways which were enriched for each geneset was organized by BH value. Each pathway was significant for up-regulated expression.

DISCUSSION

Raw Affymetrix UI33 Plus 2.0 chip data has been normalized, RLE and NUSE scores evaluated, corrected for batch effects and PCA has been performed on the cleaned data. The quality control of the RLE and NUSE scores surmise the data is of good quality, the batch effects were corrected using `ComBat`. Furthermore, Principal Component analysis was completed on the normalized data using the `prcomp()` function.

Using the methods summarized previously, the most variant and discriminant probe sets within the datasets were selected in order to perform further analysis and understand the clinical and molecular relevance of the colon cancer samples. Given we were not able to perform hierarchical clustering on our samples, further classification of the dataset into appropriate tumor subtypes would aid in identifying specific associations between the expression profiles and biological pathways. In Table 1, significantly up-regulated genes were identified including *secreted frizzled-related protein 2* (SFRP2). SFRP2 is involved in stem cell regulation and often associated with the presence of CRC upon methylation (5). Another significantly up-regulated gene is *growth arrest-specific 1* (GAS1). Given GAS1's role as a tumor suppressor gene, further investigation into the subtype-specific gene expression profiles is necessary in order to understand these findings in a clinical context with regards to disease aggression. Interestingly, upon comparison with the paper's findings, up-regulation of both of these genes is observed among only the C1 and C3 subtypes [1]. Among the down-regulated gene findings in Table 2, *Fc fragment of IgG binding protein* (FCGBP) and *leucine rich repeat containing 31* (LRRC31) are two of the only genes we obtained which were also found in the paper. Similar to the relationship between C1 and C3 with regards to up-regulation of SRFP2 and GAS1, down-regulation of FCGBP and LRRC31 in the paper share mutual expression within subtypes C3 and C6.

Following this analysis, associations were established between the genesets specified in Table 3 and the gene expression profiles through enrichment of particular biological pathways. Similar to that of the paper, pathways involved in cell communication and cell motility were up-regulated. The paper specifically reports KEGG genesets pertaining to receptor interaction and focal adhesion, which our data was able to replicate. Also observed is the enrichment of the up-regulation of Hallmark's epithelial-mesenchymal transition (EMT). This is expected given the pervasive role that the mechanism plays in tumor invasion and metastasis [7]. Relatedly, mesenchymal stem cells are generally involved within the EMT phenomenon and often secrete paracrine factors. Prior studies demonstrate the role of paracrine factors, including the previously mentioned SFRP2 protein, which can assist in this process [8]. Though a considerably meaningful biological interpretation, further analysis could be done if subtype clustering was performed.

It is also worth noting that the identified enriched genesets are sorted by p-value as adjusted to the Bejamini-Hochberg procedure. With this in mind, our results are likely attributed to the arbitrary parameters specific to a cut-off p-value of 0.05. Had this threshold been raised, the findings could have then been interpreted differently. Investigating the resulting effects in a biological context, such as with changes in enrichment of up- or de-regulation of different biological pathways, would be interesting to compare to that of the assigned paper in future projects. Though some of the results from the original paper were replicated with regards to differential gene expression and geneset enrichment, other aspects involving hierarchical clustering were not reproducible due to the lack of analysis data available. Despite

this limitation, given the relative consistency between our findings presented to that of the paper's, we are confident that we have demonstrated the functionality of our methods.

CONCLUSION

Colon cancer (CC) is a heterogeneous disease, and currently, pathological staging fails to predict the recurrence of CC in patients. Thus, this paper focuses on establishing the approach based on mRNA expression profile analyses. 750 samples were collected from various sources and were characterised into subgroups based on DNA alteration, chromosomal instability status. The screening performed utilized the whole genome and transcriptome array. As a result, 566 samples were able to pass the RNA quality requirement. Given that we were unable to perform hierarchical clustering on the gene expression data, molecular subtypes were not identified. Despite this, the provided data was processed, chip data was normalized, and RLE and NUSE scores were calculated. Both the RLE scores histogram and NUSE scores histogram reveal our data was good quality. After cleaning the data, batch effect and PCA was implemented to assist in understanding the data quality. After quality control, differential expression analysis and geneset enrichment was performed. We found meaningful insight, as the significantly up-regulated genes, SFRT2 and GAS1, were observed. These genes were seen to be associated with specific CRC subtypes, C1 and C3, upon comparison to the paper. Down-regulated genes including FCGBP and LRRC31 were some of the only down-regulated genes observed in the paper's results. This may be due to missorting of the data by adjusted p-value. Upon computing the Fisher's test and conducting genset enrichment using the KEGG, GO, and Hallmark gene sets, our results showed cell communication pathways and epithelial-mesenchymal transition processes were significantly enriched.

The biggest challenge that our group has faced is lack of analyst role participation and the expected output file. That plays a major role in understanding the data and finding meaningful interpretation from the statistics. However, we try to cover that band with the help of a sample file provided. We could reach the 50% similar results as of the research paper but it seems like we could not be able to do all the interpretations due to the gap. A future goal our group hopes to accomplish is to take time to review what each person accomplishes at regular meetings. This will help each person understand the skills and effort put in by other group members, as well as a better understanding of the project/ project goal as a whole.

REFERENCES

1. Wang, Y., Jatkoe, T., Zhang, Y., Mutch, M. G., Talantov, D., Jiang, J., . . . Atkins, D. (2004). Gene expression profiles and molecular markers to predict recurrence of dukes' b colon cancer. *Journal of Clinical Oncology, 22*(9), 1564-1571. doi:10.1200/jco.2004.08.186
2. Eschrich, S., Yang, I., Bloom, G., Kwong, K. Y., Boulware, D., Cantor, A., . . . Yeatman, T. J. (2005). Molecular staging for survival prediction of colorectal cancer patients. *Journal of Clinical Oncology, 23*(15), 3526-3535. doi:10.1200/jco.2005.00.695
3. Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., . . . Tollenaar, R. (2011). Gene expression signature to improve prognosis prediction of stage ii and iii colorectal cancer. *Journal of Clinical Oncology, 29*(1), 17-24. doi:10.1200/jco.2010.30.1077
4. O'Connell, M. J., Lavery, I., Yothers, G., Paik, S., Clark-Langone, K. M., Lopatin, M., . . . Wolmark, N. (2010). Relationship between tumor gene expression and recurrence in four

independent studies of patients with stage ii/iii colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin. *Journal of Clinical Oncology, 28*(25), 3937-3944. doi:10.1200/jco.2010.28.9538

5. Marisa, L., De Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., . . . Boige, V. (2013). Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, validation, And Prognostic Value. *PLoS Medicine, 10*(5). doi:10.1371/journal.pmed.1001453
6. MSigDB collections: Details and acknowledgments. (n.d.). Retrieved February 24, 2021, from http://www.gsea-msigdb.org/gsea/msigdb/collection_details.jsp
7. Cao, H., Xu, E., Liu, H., Wan, L., & Lai, M. (2015). Epithelial-mesenchymal transition in colorectal cancer metastasis: A system review. *Pathology, research and practice*, *211*(8), 557–569. https://doi.org/10.1016/j.prp.2015.05.010
8. Mirotsou, M., Zhang, Z., Deb, A., Zhang, L., Gnecchi, M., Noiseux, N., . . . Dzau, V. (2007). Secreted frizzled related Protein 2 (SFRP2) is the key Akt-mesenchymal stem Cell-released paracrine FACTOR mediating Myocardial survival and repair. *Proceedings of the National Academy of Sciences, 104*(5), 1643-1648. doi:10.1073/pnas.0610024104

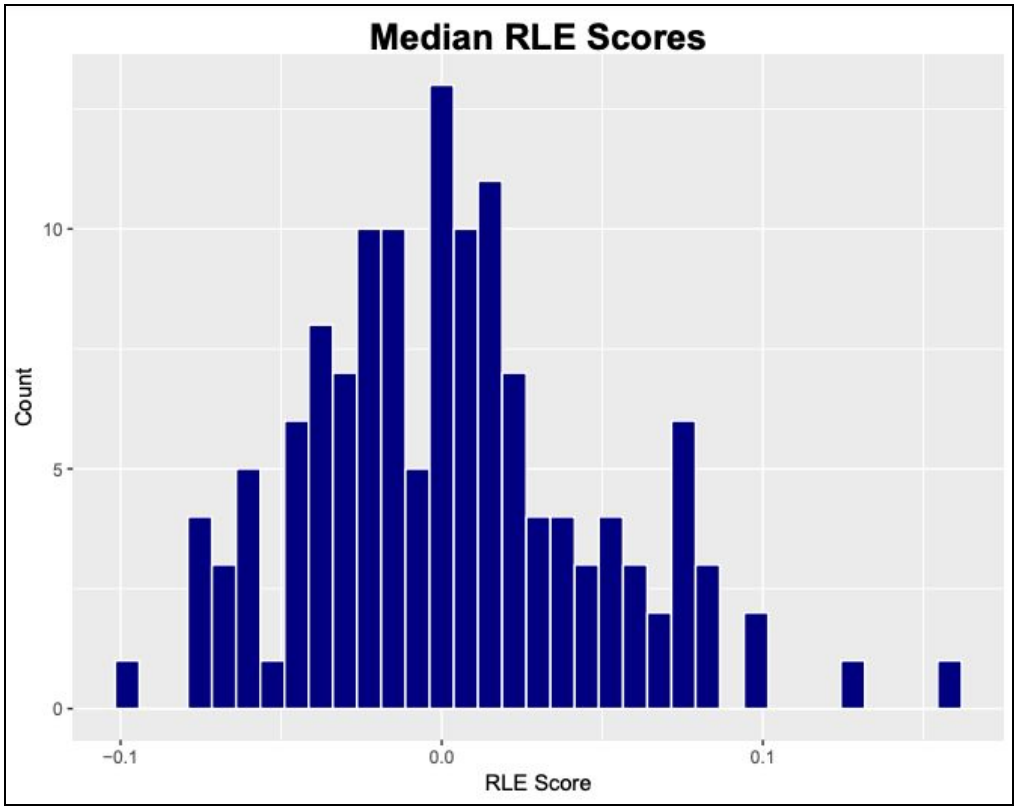


Figure 1. RLE Scores. Histogram made of Median RLE Scores for all sample data split into bins of 35, histogram shows normal distribution of RLE scores centered over 0 indicating good quality data.

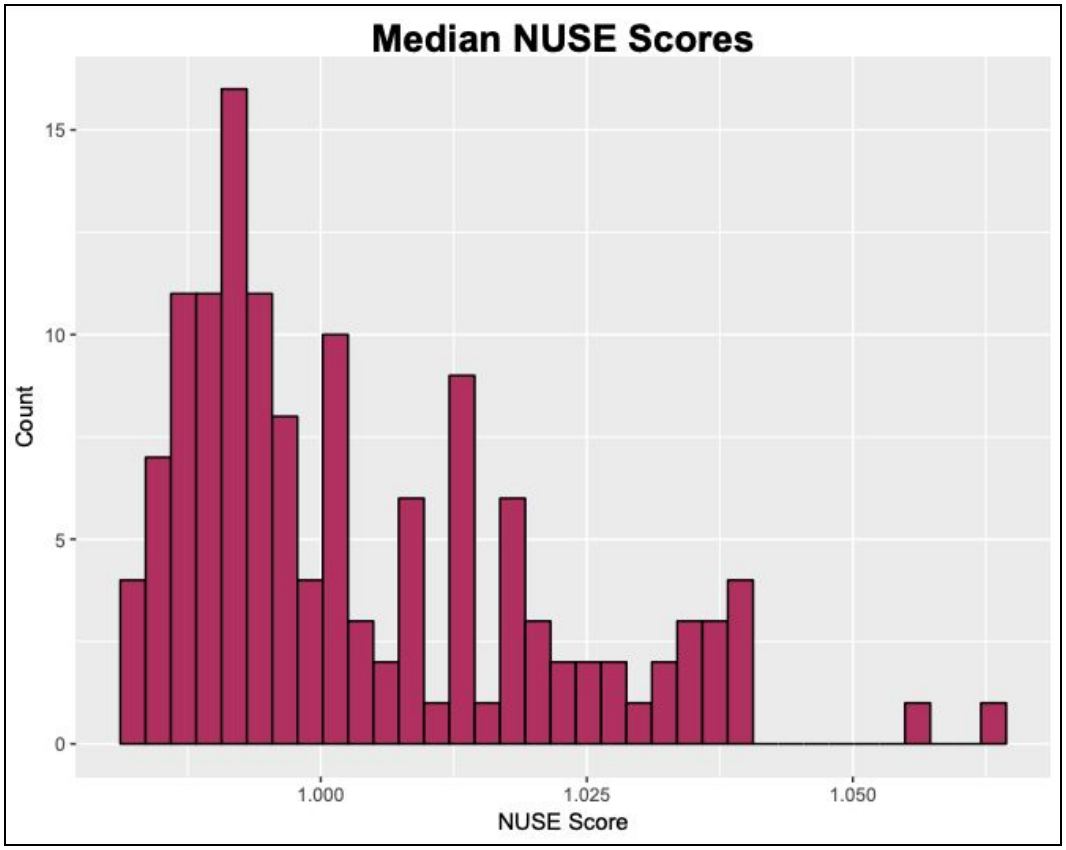


Figure 2. NUSE Scores. Histogram made of Median NUSE Scores for all sample data split into bins of 35, histogram shows skewed distribution of NUSE scores mostly less than 1, indicating good quality data.

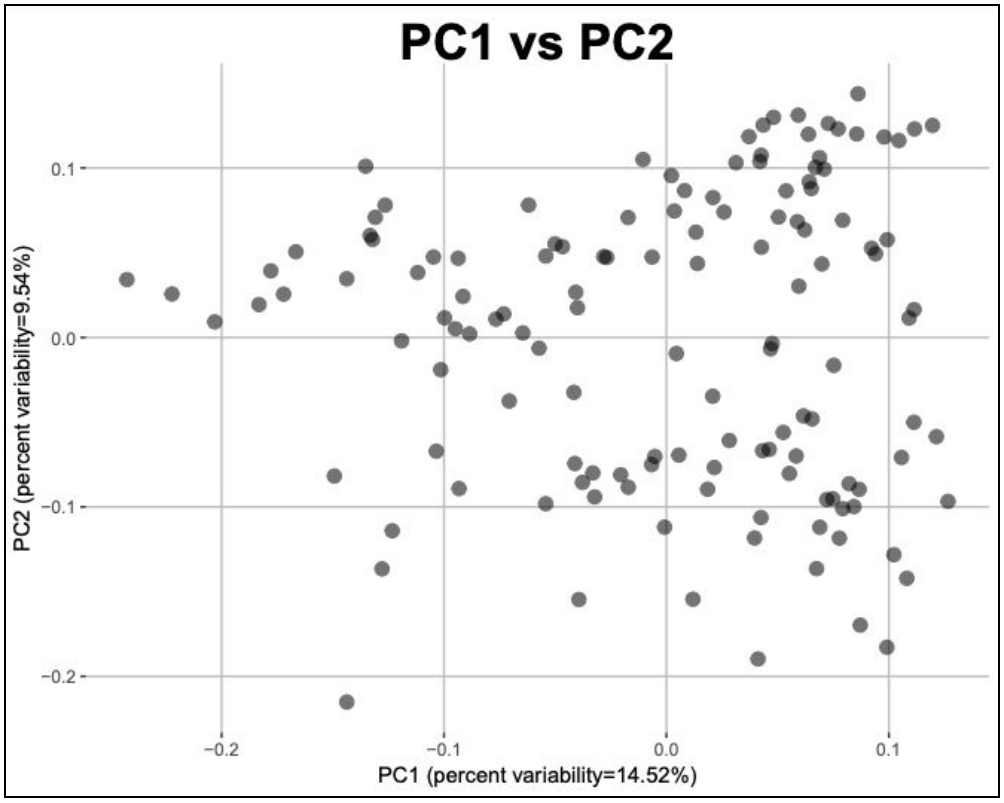
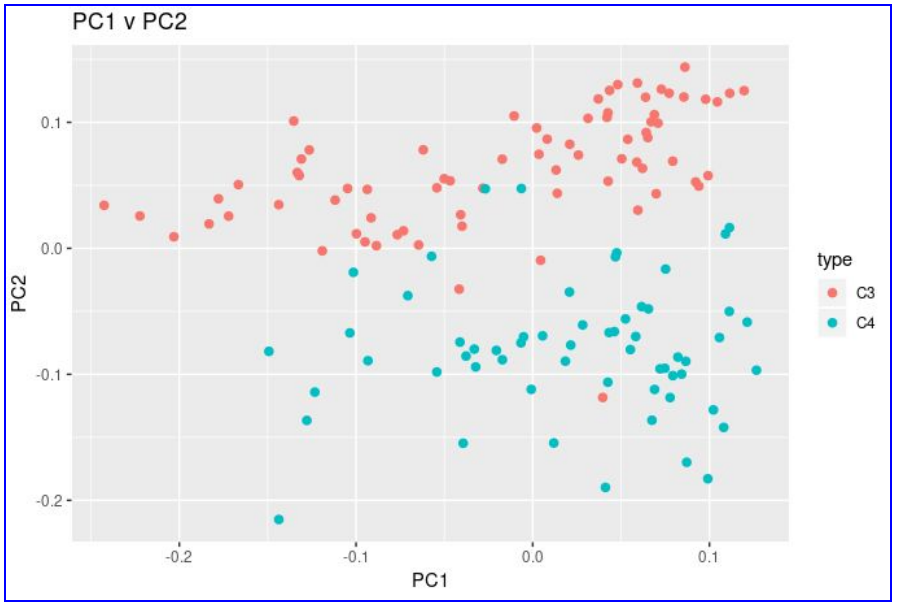


Figure 3. PC1 v PC2. Dot plot of PC1 vs PC2 with percent variability of each PC on the relevant axis.



(**Supplementary Figure 1. PC1 v PC2 colored by C3 and C4.** Dot plot from Figure 3 colored by C3 or C4 types indicating clustering of each type (not produced by this group).