Project: Microarray-based Tumor Classification

Divya Venkatraman, Garima Lohani, Marlene Tejeda, Xudong Han
(Biologist) (Data curator) (Programmer) (Analyst)
Group 1
TA : Kritika Karri

INTRODUCTION:

Colon Cancer (CC) is one of the most common causes of human death by cancer. There is no reliable method to detect the recurrence of colon cancer in patients. There is a lack of availability of reliable molecular markers to identify the tumor because of its heterogeneous nature [1]. However, microsatellite instability (MSI) has only been found reproducible to detect early colon cancer but technology needs to be improved and refined [3].Microarray in recent years has generated a lot of gene expression profile (GEP) of CC. In the recent studies, GEP studies that include unsupervised hierarchical clustering have found at least three distinct molecular subtypes of CC. Therefore, microarray gene expression profile is one of the technologies that can help us get the standard and reproducible molecular classification [1].In this project, we are trying to identify and compare C3 and C4 subtype of CC using hierarchical clustering, which is an unsupervised machine learning algorithm.

DATA:

The data consists of gene expression of human colon cancer (CC) cells. They were collected from Cartes d'Identité des Tumeurs (CIT) program from the French Ligue Nationale Contre le Cancer. In this program, the fresh frozen primary tumor cells were collected from a cohort of 750 patients. They were categorized from stage I to stage IV according to the American Joint Committee on Cancer tumor node metastasis (TNM) staging system. Out of 750 tumor samples, 566 passed RNA quality requirements. These samples were divided into discovery set (n=443) and validation set (n=123). Also, the validation set included the 906 public datasets. There were 19 non tumoral dataset. The data also recorded if there was any presence of adjuvant chemotherapy in patients from stage II to III CC. The study investigated the seven most frequent mutations in codon 12 and 13 of KRAS. In addition, BRAF and TP53 gene mutations were also assessed. MSI tumors were analyzed and were divided into deficient MMR (dMMR) and proficient MMR (pMMR). CIMP (CpG island methylator phenotype) status was based on five biomarkers (CACNA1G, IGF2, NEUROG1, RUNX3, and SOCS1) [1].

The protocol for RNA extraction included treatment with liquid nitrogen. Then RNAble was used for the extraction of RNA. The extracted RNA was placed on RNAeasy columns. Then there was analysis with electrophoresis and its quantification using Nanodrop ND-1000. After Microarray analysis was performed, RNA was labeled as per the manufacturer's one-cycle target labeling protocol. After hybridization to HG-U133 Plus 2.0 Affymetrix GeneChip array, the chips were scanned with a GCOS 1.4. Microarray data [2]. The data is accessible at NCBI GEO (accession number: GSE39582).The author identified six molecular subtypes of CC in this paper. But our analysis is focused on reproducing results for C3 and C4 subtypes.

METHODS:

Symbolic links

There were 133 CEL files at a location (/project/bf528/project_1/data/GSE39582/CEL_files) on SCC. For these samples, the symbolic links were created to remove redundancy and for the reusability of these files. GSM971958, one of the files was downloaded from NCBI GEO. So, there were a total of 134 CEL files for further analysis.

Reading In Files

With the affy library, we read in the different CEL files in R using the function ReadAffy. This function allows multiple files to be read in simultaneously. ReadAffy function is quite flexible as it allows users to specify file type and phenotype information [4].

Normalization of Data

Afterwards, all the files were normalized using the rma function in R. RMA function consist of three steps [5]:

1. Background correction: It adjusts raw PM probe intensities using a model. Additionally, it corrects for background noise and preprocessing effects and adjusts for cross hybridization of non-specific DNA to the array.
2. Quantile Normalization: Process of removing unwanted non-biological variation between the chips in the microarray datasets. Method gives the same distribution of intensities to each array.
3. Summarization: Combines probe intensities across arrays.

Quality Assurance

Bioconductor package AffyPLM was used to compute the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores in the datasets. Package AffyPLM provides quality assessment on given datasets to check for instances such as RNA degradation, as a result, RLE and NUSE, was used on un-normalized files. NUSE estimates the standard error for each array from the PLM fit [6] which normalized and removes background noise. This method checks for variability between genes by adjusting the standard error so that the median across the arrays is 1 for each gene. RLE computes the estimates of expression on a log scale for each gene on each array and then computes the median value across arrays for each gene [6]. Arrays that are problematic are not centered at zero or have large spread. To access this histograms were created to visualize the distribution of RLE and NUSE scores.

Correct for Batch Effects

Sva package was used to correct batch effects and other unwanted variations using the ComBat function using known batches[7]. Batch variables are passed as separate arguments and the output is the removal of batch effects.

Principal Component Analysis

PCA is widely used to reduce the dimensionality of the data and allow for summarization of the data. PCA function was used on the normalized data using the prcomp[8] function and scaled and center. The data was transposed to be able to scale within each gene rather than sample. Outliers were examined by examining the boxplot of the first two principal components that were three standard deviations away from the mean.

Noise Filtering

To remove the genes without significant variance among 134 samples in the expression level, we performed the noise filtering analysis against the data (Table S1) in three steps. Genes expressed in under 20% of samples were removed first. Next, we eliminated genes with variances not significantly different from the median variance of all probe sets via a chi-squared test using a threshold $P > 0.01$ (Table S2). In the last, we filtered out the genes whose coefficients of variation are less than 0.186 (Table S3).

Clustering Analysis and Subtype Discovery

We used the hierarchical clustering method to separate the genes in Table S3 into two groups based on the dendrogram. According to the dendrogram, the height 100 was chosen to cut the tree into two groups. To identify the differentially expressed genes in two clusters, we performed a students' t-test to each gene and selected the genes with adjusted $p < 0.05$ (Table S4). The top 10 genes with the biggest t-statistic were chosen as the most upregulated genes in the C4 subtype compared to C3, and vice versa.

Gene Set Enrichment

We use the genes that passed the three screening tests to perform functional enrichment. Using the Hgu133plus2.db package in Bioconductor, we mapped the gene symbols to the probeset IDs. Some symbols map to multiple probeset IDs. In such cases, we chose the probeset ID with the highest fold change value and discarded the rest. The genes with highest significance were chosen because in further downstream analysis, it is important to have the most significant genes in our data. We selected the top 1000 up regulated and down regulated genes using the highest and lowest t-statistic values.

We downloaded the KEGG, GO and Hallmark gene set collections from MsigDB[10] and loaded it in R using the GSEABase package in Bioconductor[11], to use for our gene set enrichment. The gene lists of top 1000 differentially regulated genes were enriched using the gene sets in each collection by performing fisher exact test. We adjusted the p-value of the fisher exact test for multiple hypotheses using Benjamini-Hochberg method[12].

Implementation

All computing procedures were coded in the R language in R v3.5.1, which is widely used for statistical programming.The AFFY v1.64 and affyPLM v 1.62.0 packages in Bioconductor v1.6 were used. The SVA package v3.34.0 was used for batch effect removal. The GSEABase package in Bioconductor was used to download the gmt files. Hgu133plus2.db package in Bioconductor was used for annotation. Overall the runtime for each of the analysis was very quick.

RESULTS:

Quality Control for the Microarray Expression Data Identify Three Outliers

We performed a quality control analysis to examine the reliability of the microarray data. After normalization and batch effect correction, we computed the median of RLE and NUSE for each chip based on the PLM and summarized them in the histogram (Fig1). Most of the RLE medians are very close to 0 (mean 0.00265, standard deviation 0.045) (Fig1A). Furthermore, most of the NUSE medians are close to or only slightly bigger than 1 (Fig1B). These pieces of evidence indicate that most of the chips we utilized in this analysis are of good quality. Subsequently, a PCA analysis was conducted to identify the chips with a higher tendency to be the outliers (Fig2). We could not easily find clusters based on our PCA1 vs. PCA2 plot (Fig2A). However, we discovered three outliers when we built the box plots to these first two principle components individually (Fig3A). These three outliers in our dataset are corresponding to the GSM972097.05-0805.04.CEL.gz, GSM972350_MFL_036b_U133_2.CEL.gz, and GSM972467_MFL_400-b_U133_2.CEL.gz files. Two of these are C4 samples, and the other is the C3 sample.

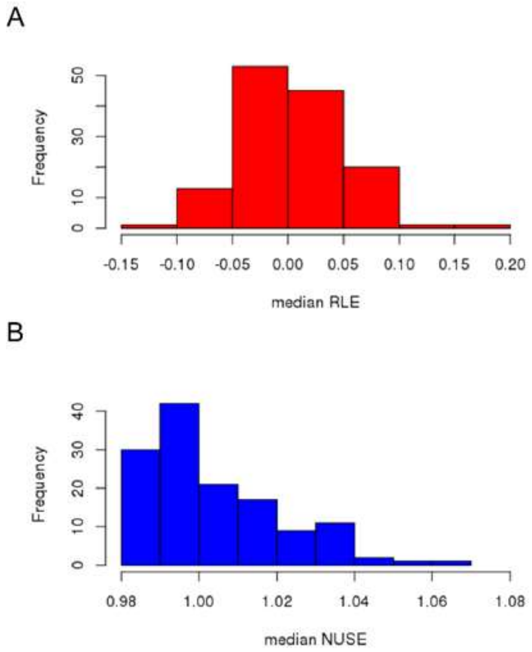

Fig1. RLE and NUSE median distribution histograms based on PLM after background adjustment and normalization. (A) The histogram of median RLE for all the samples. The median larger than 0 indicates more genes expression in the sample are up regulated. (B) The histogram of median NUSE for all samples. The median bigger than 1 indicates the chip may have poor quality.

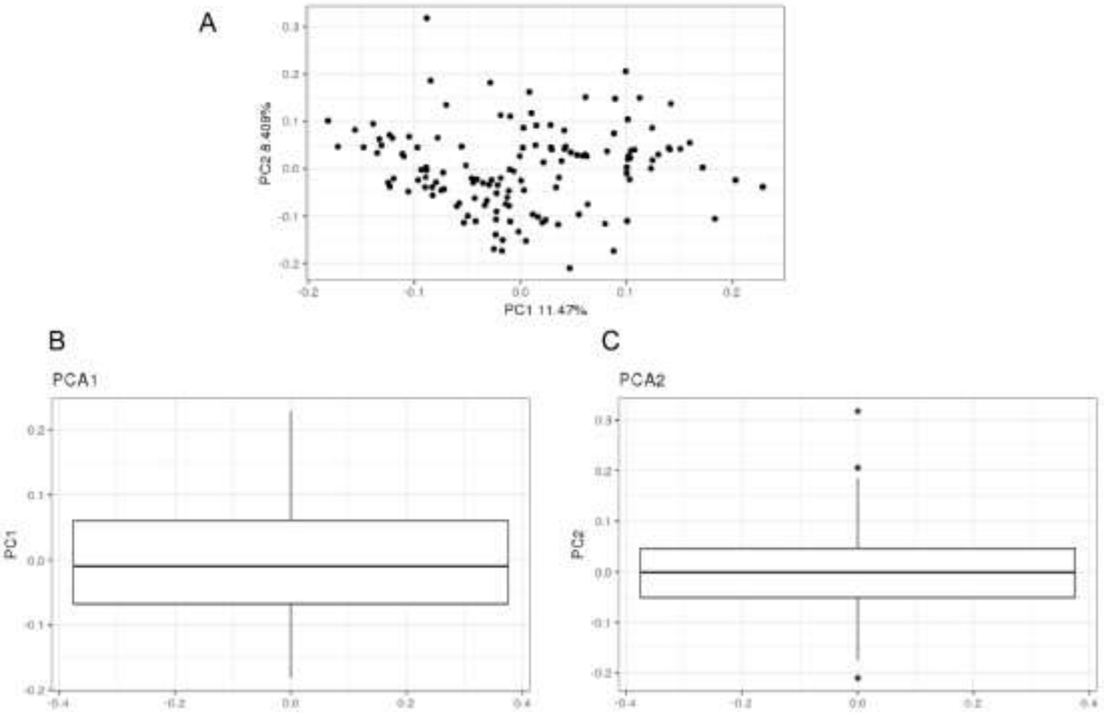

Fig2. PCA1 vs PCA2 with its respective percent variance. (A) Scatter plot shows that there is no evident clustering between the two PCs. PC1 and PC2 have 11.47% and 8.4% variance respectively. (B) (C) Boxplot for PC1 and PC2. Figure C shows that there are three outliers in PCA2 data. 2 of 3 are C4 subtype, and another is C3 subtype.

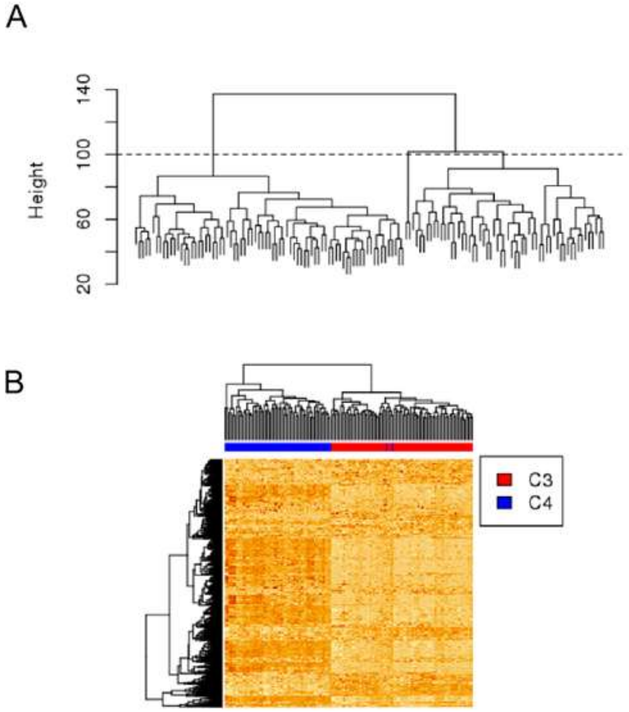

Fig3. The clustering analysis. (A) Hierarchical clustering of the samples based on the expression 1531 filtered genes (see methods). The dashed line is the height to cut the tree into two groups. There are 57 samples in the cluster1 (left) and 77 samples in the cluster2 (right). (B) The heatmap of the filtered genes. The red and blue colors in the top bar denote the C3 (75) and C4 (59) subtype samples respectively. All the samples in the cluster1 (57) are C4 subtype. All the C3 (75) samples and only two remaining C4 samples are in cluster2.

Unsupervised Analysis of Filtered Gene Expression Profiles Separated Samples into Two Groups

To select the most representative genes that can distinguish the C3/C4 colon cancer subtypes, we used three screening methods (see methods) to keep the genes passing all filters (TableS4). We obtained 39661 genes with expression intensity in at least 20% samples from all 54675 genes, and we obtained 15508 genes with significant variation across all samples. Finally, after the third filter, we kept 1531 genes that may be the signature to classify C3 and C4 cancer subtype.

To classify these samples based on the selected genes above, we utilized an unsupervised hierarchical clustering analysis against 134 samples (Fig 2). The dendrogram (Fig 2A) showed that these samples could be separated into two clusters (57 in cluster1, 77 in cluster 2). Based on the heatmap (Fig 2B), cluster1 and cluster2 can respectively match to C4 and C3 subtypes very well except for two C4 samples (GSM972019_080705.02.CEL.gz, GSM972412_VB_067T_U133_2_2.CEL.gz) located in cluster2 (in fisher test).

Furthermore, we performed a Welch t-test to identify the genes that may significantly differentially express in these two clusters, and we obtained 1236 genes with a q value of less than 0.05. These genes may be the signature to predict and classify the subtype of colon cancers. The most upregulated genes in the C3 samples compared to C4 samples are FCGBP, ST6GalNAcI, MUC2 etc., and the most downregulated genes are RBMS1, CCDC80, Hs.65029 etc.

Biological Relevance of Enriched Gene Sets

We used the 3 gene set collections from MsigDB [10] to enrich the top 1000 up and down regulated genes. The top 10 upregulated and downregulated genes are shown in Table 1 and Table 2. The Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. This gene set collection contains 50 gene sets. The Kyoto Encyclopedia of Genes and Genomics (KEGG) gene set collection has canonical pathways representing knowledge on the molecular interaction, reaction and relation networks of different biological processes. This collection has 186 gene sets. The Gene Ontology (GO) collection has gene sets annotated by the same GO term. This collection has 9996 gene sets. We obtain 540 gene sets enriched with p-value < 0.05.

The top 3 significantly enriched gene sets in the GO collection are given in Table 3. These gene sets were enriched by the down regulated gene list. The gene sets enriched include protein complex binding, which is proteins that interact selectively and non-covalently with a macromolecular complex [10]. Glycosaminoglycan binding gene set which has proteins that interact selectively and non-covalently with any glycan (polysaccharide) containing a substantial proportion of aminomonosaccharide residues [10] was also significantly enriched. The negative regulation of cell differentiation gene set has proteins that stops, prevents, or reduces the frequency, rate or extent of cell differentiation [10].

In the Hallmark collection, epithelial mesenchymal transition which have genes defining wound healing, fibrosis and metastasis was found to be upregulated as well as downregulated. UV response gene set which have genes down-regulated in response to ultraviolet (UV) radiation [10], was enriched by the down regulated genes as shown in Table 4. The fatty acid metabolism gene set which has proteins involved in metabolism of fatty acids was enriched by the upregulated genes.

The three gene sets enriched in the KEGG collection were all enriched by up regulated genes as shown in Table 5. These three enriched gene sets are related to metabolism of drugs and butanoate.

Gene Symbol	**t-statistic**	**p-value**	**Adjusted p-value**
GAS1	24.2618708130683	4.43119620847574E-49	6.78416139517635E-46
CCDC80	22.6880916598421	3.10349184535399E-46	1.18786150380924E-43
RBMS1	22.6626455161317	3.84404660214157E-47	2.94261767393937E-44
ARMCX1	22.5141581868672	4.985648028944E-45	1.09043244747332E-42
SFRP2	22.3516576715084	1.06601997688104E-46	5.44025528201626E-44
GPC6	22.1272130560437	4.68308051906206E-45	1.09043244747332E-42
MGP	22.0930982486816	9.0941651562712E-46	2.78463337085024E-43
FNDC1	21.7407554259285	6.47508724604741E-45	1.10148428596651E-42
MSRB3	21.5985719144549	6.36289990256921E-45	1.10148428596651E-42
SPOCK1	21.5116575677259	3.27631035329937E-44	5.01603115090133E-42

Table 1. Top 10 up regulated genes based on t-statistic value

Gene Symbol	**t-statistic**	**p-value**	**Adjusted p-value**
FCGBP	-16.2191139075554	4.6907054739414E-29	3.98970560033571E-28
ST6GALNAC1	-13.4923395469676	2.82044436544213E-23	1.43458482508037E-22
MUC2	-13.4069664246303	1.76587962083503E-24	9.79551340397981E-24
CLCA1	-13.2744461407883	2.9029608858388E-25	1.70284793724874E-24
LRRC31	-13.1346067365202	1.469233470189E-25	8.99758577143741E-25
CCL28	-12.7252878640803	1.08605591216437E-24	6.11305735854284E-24
HEPACAM2	-12.5863443757738	5.03317932898199E-24	2.67562415023313E-23

NXPE1	-12.5186017055264	8.61718481495261E-24	4.47217286498049E-23
DUOXA2	-12.292619764974	5.88602422082973E-23	2.89758941546312E-22
SPINK4	-12.2122467093841	7.67065158413572E-22	3.43385016821982E-21

Table 2. Top 10 downregulated genes based on t-statistic value

GO Gene Set Name	**p-value**	**estimate**	**BH adjusted p-value**
GO_PROTEIN_CONTAINING_COMPLEX_BINDING	1.093E-05	0.276	0.218
GO_GLYCOSAMINOGLYCAN_BINDING	4.320E-05	0.256	0.235
GO_NEGATIVE_REGULATION_OF_CELL_DIFFERENTIATION	4.838E-05	0.303	0.235

Table 3. Enriched gene sets in GO collection based on p-value

Hallmark Gene Set Name	**p-value**	**estimate**	**BH adjusted p-value**
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	3.485E-07	0.220	3.485E-05
HALLMARK_UV_RESPONSE_DN	0.005	0.244	0.188
HALLMARK_FATTY_ACID_METABOLISM	0.011	0.207	0.284

Table 4. Enriched gene sets in Hallmark collection based on p-value

KEGG Gene Set Name	**p-value**	**estimate**	**BH adjusted p-value**
KEGG_DRUG_METABOLISM_CYTOCHROME_P450	0.0011	0.0919	0.414
KEGG_BUTANOATE_METABOLISM	0.0099	0.077	0.596
KEGG_DRUG_METABOLISM_OTHER_ENZYMES	0.0099	0.077	0.596

Table 5. Enriched gene sets in KEGG collection based on p-value

DISCUSSIONS:

In order to conduct the analysis all 134 CEL files were read in using ReadAffy function and then the data was normalized. To validate the quality and reliability of the microarray data AffyPLM was used to compute the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores in the data. Furthermore, after correcting for batch effects, PCA was run to identify the chips with a higher tendency to be the outliers. Nose filtering was conducted to retain the genes with the most differentially expression. We then did an unsupervised hierarchical clustering to classify all the samples. Finally, We performed gene set enrichment to identify genes that are over-represented in a large set of genes.

Box plots and histograms of probe set intensities levels across microarrays are widely used methods to access data homogeneity. Methods like RLE and NUSE are used to visualize outliers in a group of microarrays [9]. RLE measures the expression of a particular probe set in a chip and its deviation from other chips. It is assumed that the majority of genes are not biologically affected and the number of regulated genes roughly equal the number of down regulated genes[9] . For this reason, the median of RLE from zero is a measure of variability in a given chip. Ideally, the median should be around zero, which is what was demonstrated for this dataset (Figure 1). NUSE is based on a robust linear model fit to each probe set and because probes within a set are supposed to be consistent in terms of expression [9]. NUSE measures the relative precision of expression estimates. A good quality expression array, the median NUSE would be around one. Figure 1 illustrates a median value for NUSE to be around one, indicating a good expression array.

Principal component analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the dataset and helps identify any clustering between the genes. Figure 2 shows that the first principal component in the x axis has the largest variation; the second principal component shows the direction uncorrelated to the first principal component. The first two components explain roughly 20% of the variance. There isn't any clear clustering between the genes (Figure 2). Furthermore, to evaluate any outliers that were three standard deviations away from the mean, boxplots were created for the first two PCs. As can be seen in Figure 3, the first principal component shows no outliers, however, the second principal component shows one outlier three deviations away from the mean. Given the large scale of the dataset, this point was not removed.

The noise filtering can help us delete the bulk of genes that do not express differentially across all the samples and retain the representative genes as the potential signature to distinguish these cancer subtypes. Combining these filtered data with unsupervised clustering method can further enable us to classify these samples into two clusters corresponding to the C3 and C4 subtypes. The clustering results in our project show a high consistency with the results from the reference [1].

The gene set enrichment process gives us the biological relevance of our results. The top 10 deregulated genes that we found include SFRP2, which is key to stem cell regulation and GAS1 *growth arrest-specific 1*, were mentioned in the paper as probable markers of aggressiveness of CC cells [1]. The gene sets that represent metabolism of cytochrome P450 and butanoate and epithelial mesenchymal transition were shown to be enriched in the reference paper [1] . The reference paper shows that epithelial mesenchymal transition is downregulated in C4 and upregulated in C3 which is probably the reason why we have found it as both up and down regulated. The other gene sets that we found as significantly down regulated were related to protein binding and regulation of cell differentiation which may be a cause for rapid progression of metastatic carcinomas.

CONCLUSIONS:

In this project, we analyzed the microarray data for 134 colon cancer samples. We identified three chips as possible outliers and selected 1236 genes as the potential signature to classify or predict the C3 and C4 cancer subtypes. Gene set enrichment of the top 1000 deregulated genes from the selected genes show that metabolic pathways get upregulated and pathways related to cell differentiation and protein binding are downregulated which may contribute to cancer progression and metastasis.

REFERENCES:

[1] Marisa, L., Reyniès, A. D., 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).

[2] de Reyniès A, Assié G, Rickman DS, Tissier F, Groussin L, et al. (2009) Gene expression profiling reveals a new classification of adrenocortical tumors and identifies molecular predictors of malignancy and survival. J Clin Oncol 27: 1108–1115.

[3] Popat S, Hubner R, Houlston RS (2005) Systematic review of microsatellite instability and colorectal cancer prognosis. J Clin Oncol 23: 609–618.

[4]Gautier, Laurent, Rafael Irizarry, Leslie Cope, and Ben Bolstad. 2013. “Description of Affy.” https://www.bioconductor.org/packages/devel/bioc/vignettes/affy/inst/doc/affy.pdf.

[5] Chen, Ding-Geng (din), Karl E. Peace, and Pinggao Zhang. 2017. *Clinical Trial Data Analysis Using R and SAS*. CRC Press.

[6] Bolstad, Ben. 2011. “affyPLM: Model Based QC Assessment of Affymetrix GeneChips.” https://rdrr.io/bioc/affyPLM/f/inst/doc/QualityAssess.pdf.

[7] Leek, Jeffrey T., W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. 2012. “The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments.” *Bioinformatics* 28 (6): 882–83.

[8] Tang, Hui, and Terry M. Therneau. 2010. “Statistical Metrics for Quality Assessment of High-Density Tiling Array Data.” *Biometrics* 66 (2): 630–35.

[9] Wang, Antai, and Edmund A. Gehan. 2005. “Gene Selection for Microarray Data Analysis Using Principal Component Analysis.” *Statistics in Medicine* 24 (13): 2069–87.

[10] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, Jill P. Mesirov. 2011. “Molecular signatures database (MSigDB) 3.0”. *Bioinformatics* 27(12): 1739–1740

[11] Morgan M, Falcon S, Gentleman R. 2019. GSEABase: Gene set enrichment data structures and methods. R package version 1.48.0

[12] Benjamini, Yoav; Hochberg, Yosef. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing" . Journal of the Royal Statistical Society, Series B. 57(1): 289–300