|  |
| --- |
| Replicating machine learning analysis of gene expression profiles in cervical cancer  Rebecca Lindner1  1Data Science Institute, Columbia University  Abstract  **Motivation:** Cervical cancer is the fourth most common cancer among women across the world. A previous study proposed an integrative machine learning approach to analyzing multiple gene expression profiles to identify genetic markers associated with cervical cancer, which was replicated here. Such a set of genetic markers could be used in the diagnosis and treatment of cervical cancer in the future.  **Results:** Replicating the steps performed in this study proved challenging based on the limited information provided in the paper on specific methods used, and different gene sets than those noted in the original study were found through gene expression analysis of four datasets. Meta-analysis produced similar up- and down-regulated gene sets when performed on the meta-gene set from the paper, but only one overlapping gene was found for meta-analysis performed with the meta-gene set obtained here. Hierarchical clustering performed similarly.  **Availability:** Source code, implemented in R, available at <https://github.com/rlindner27/genomic-info>  **Contact:** rml2183@columbia.edu |

# Introduction

Cervical cancer is the fourth most common cancer in women, with approximately 570,000 new cases in 2018 alone, representing 6.6% of all female cancers (World Health Organization, 2018). Human Papilloma Virus (HPV) plays a primary role in the development of cervical cancer, with studies finding HPV in over 90% of cervical cancer samples (F. Xavier Bosch, 1995). While HPV is the most important risk factor for cervical cancer, others include smoking, immunosuppression, chlamydia, obesity, long-term contraceptive and intra-uterine device (IUD) use, and family history of cervical cancer (American Cancer Society, 2018). Genetic pathways may also play a role in the development of cervical cancer, and a small number of studies have looked at this through gene expression profiling (E L Ivansson, 2011), (Kejia Wu, 2018), (Wong YF, 2006).

One study in particular carried out a unique meta-analysis model in analyzing four microarray datasets (Mei Sze Tan, 2018). The process includes: (i) gene expression analysis of each individual dataset, (ii) integrating datasets to perform meta-analysis, and (iii) machine learning analysis. This paper attempts to replicate the steps performed in that study in order to assess the method outlined. Findings are reported both for the analysis performed and for comparison to the findings of the initial work.

# Methods

All analysis was performed using R (R Core Team, 2018). The four datasets were obtained from the National Center for Biotechnology Information Gene Expression Omnibus ([NCBI GEO](http://www.ncbi.nlm.nih.gov/geo)) using the package *GEOquery* (Davis, 2007). [GSE26511](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26511) included 39 samples of lymph nodes in early-stage cervical cancer patients, 20 with negative lymph node metastases and 19 with positive (Noordhuis MG, 2011). [GSE5787](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5787) included 33 samples of cervical cancer biopsies from 11 patients, with up to seven biopsies per patient to examine intratumor heterogeneity (Bachtiary B, 2006). [GSE9750](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9750) included 66 samples, 33 of cervical cancer tumors, 9 of cell lines, and 24 of normal cervical epithelium (Scotto L, 2008). [GSE63514](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63514) included a total of 128 samples, 28 of cervical cancer, 14 of CIN 1 lesions, 22 of CIN 2 lesions, 40 of CIN3 lesions, and 24 of normal cervical epithelium (Den Boon JA, 2015). Datasets were compiled for analysis with the package *simpleaffy* (Miller, 2018). Normalization within each dataset was performed using robust multi-array average (RMA), implemented in *affy* (Gautier, 2004).

The first step performed was to carry out individual microarray analysis of each dataset. Analysis of individual gene expression datasets performed using Linear Models for Microarray Analysis (*limma*) (Ritchie, 2015). The top 10,000 differentially expressed genes in each dataset were then selected and compared to the top 10,000 genes selected in the original study. As the contrast matrix and other structure variables for LIMMA were not specified in the original work, a number of attempts was made with each dataset to obtain the most similar results possible.

The top 10,000 differentially expressed genes across all four datasets were then intersected to create a single meta-dataset for analysis using *RankProd* (Del Carratore, 2017). As GSE9750 was based on a different platform (Affymetrix Human Genome U133A Array [HG-U133A]) than the other three datasets (Affymetrix Human Genome U133 Plus 2.0 Array [HG-U133\_Plus\_2]), quantile normalization was attempted on the meta-dataset using *preprocessCore* (Bolstad, 2018). As this led to lower quality results, it was abandoned in favor if RMA-normalization alone. *RankProd* analysis was then conducted both on the intersected dataset obtained through the LIMMA analysis done here as well as on the meta-dataset provided as an appendix in the original study for consistency. These results were then compared, and the top upregulated and downregulated genes found through *RankProd* fed into a machine learning framework.

Due to poor documentation of methods used for the majority of the machine learning methods utilized in the original study, only hierarchical clustering could be replicated here. Functional pathway analysis was not conducted due to time constraints. Hierarchical clustering was performed on the gene expression values obtained through LIMMA analysis for the genes found to be differentially expressed in the original study for maximum consistency of. This clustering was then compared both through trees obtained and heatmaps produced using *gplots* (Warnes, 2016).

# Results

The objective of the results was twofold: to obtain differentially expressed genes, and to compare results to those found in the original study. Results for each step of the process are outlined below.

## Gene expression analysis of individual datasets

Varying results for each dataset were found in terms of consistency with the original study, outlined in Table 1. The number of genes found in the LIMMA analysis performed that were in the original set of 10,000 provided as an appendix to the original paper ranged from 2,055 to 5,276 depending on the dataset. A full comparison of all results obtained here and given in the original is provided in Appendix 1. While the paper stated that a quantile algorithm was applied to normalize the data prior to this step, and that all data were log2-transformed, performing these steps proved to give results even more disparate from those found in the original work, and as such this step was excluded. RMA normalization was performed on each dataset individually.

**Table 1.**Comparison of results of study findings and analysis findings with respect to LIMMA analysis of individual datasets

|  |  |  |  |
| --- | --- | --- | --- |
| Gene Dataset | Shared Genes  (top 10,000) | Shared Genes  (top 1,000) | Shared Genes  (top 100) |
| GSE26511 | 3799 | 552 | 54 |
| GSE5787 | 2055 | 186 | 18 |
| GSE9750 | 5276 | 552 | 38 |
| GSE63514 | 2556 | 290 | 37 |

Number of genes shared in top 10,000, 1,000, and 100 genes found in the study by the top 10,000 genes selected in replicated LIMMA analysis.

### 3.1.1  GSE26511

GSE26511 was found to be among the easiest datasets to replicate LIMMA analysis for, though it did not prove possible to obtain the same results consistently. LIMMA found the same average expression values when the contrast matrix was structured for positive vs. negative lymph node samples for all genes that were shared between the original top 10,000 and the found top 10,000. However, only 3,799 of the genes in the original top 10,000 were found in this analysis, and p-values differed across all genes on the order of 10-2 to 10-5. Additionally, while the original paper stated that genes were filtered for an adjusted p-value of <0.05, the appendix provided showed no genes in this dataset with an adjusted p-value below 0.18, an inconsistent documentation issue that complicated this work. 5 genes were found to have an adjusted p-value below this threshold in the replicated analysis. Several attempts at normalization and log-transformation were made with limited success.

### 3.1.2  GSE5787

The lowest accuracy was found in replicating the analysis performed on GSE5787, likely due to the many possible levels used in the contrast matrix. This dataset contained up to seven biopsies on eleven tumors, and the results most consistent with the original study were found when constructing the contrast matrix on the biopsies rather than the tumors/patients. A number of different permutations were tried, with the best results found for setting biopsy 1 as the base level and adding together and averaging the other levels for contrast. However, this still found around 20% of the top 10,000, 1,000, and 100 genes stated in the paper. As with the other datasets, normalization and log-transformation were also tried. Similarly to GSE26511, the adjusted p-values in the original paper were never lower than 0.19, despite a note that filtering was performed for <0.05. No genes were found with an adjusted p-value below this threshold in the replicated analysis.

### 3.1.3  GSE9750

GSE9750 showed fairly good performance in replicating the analysis, with the highest number of genes in the top 10,000 in the original study found (5,276). While the paper stated that normal samples were removed prior to this analysis, removing the normal samples proved to make the results less consistent. As such, the contrast matrix was structured with cancer as the base and the cell line and normal samples as the contrast. Again, normalization and log-transformation proved ineffective. 44 genes in the original study were shown in the appendix to have an adjusted p-value of <0.05, and 17 of these were found by the replicated analysis, 7 with the found adjusted p-value still <0.05. A total of 3,224 genes were found in this analysis to have an adjusted p-value below this threshold, perhaps demonstrating an issue with the normalization or contrast matrix construction used for LIMMA.

### 3.1.4  GSE63514

Finally, results for GSE63514 showed poor consistency with the original study. After several permutations, the contrast matrix producing the most similar results was found to be one using cancer as the base, with CIN 1, CIN 2, CIN 3, and normal samples summed and averaged for comparison. This still gave only 2,556 shared genes. Normalization and log-transformation were implemented with poorer results found than for the RMA-normalized data alone. As with GSE26511 and GSE5787, no genes in the original appendix had an adjusted p-value of <0.05. 9,678 genes in the analysis performed were found to have an adjusted p-value below this threshold, again demonstrating a possible issue with normalization or LIMMA construction.

## Meta-analysis of multiple datasets

As different top 10,000 genes were found for each dataset in this analysis, the intersection of these four sets was not equivalent to the meta-dataset produced in Appendix S2 of the original paper. 526 genes were found to be shared between the top 10,000 genes for GSE26511, GSE5787, GSE9750, and GSE63514 here, and 117 in the original analysis. Between these two sets, only 3 genes were shared. Due to these consistency issues, three rounds of RankProd analysis were performed with the cancer samples: one using the dataset provided by the paper in Appendix S2, one using the genes given in the paper’s dataset but incorporating the gene expression values found in this analysis, and one using a random sample of 97 genes found in the LIMMA intersection analysis along with the 3 shared genes. Sampling was performed due to the computational limitations of RankProd, as a set of over 500 genes could not be used. RankProd was performed with a p-value cut-off of <0.05 and a percentage of false positive predictions (pfp) cut-off of <0.05. Detailed results can be found in Appendix 2.

### 3.2.1  Replication of analysis with original dataset

Unsurprisingly, nearly identical results were found when performing RankProd analysis on the meta-dataset provided as an appendix to the original paper. For the up-regulated genes in particular, the fold change (FC) values differed, though all other indicators were identical. Attempts were made to log-transform the data and the resulting FC value with various base systems that did not prove successful in creating comparable FC values. For the up-regulated genes (Table 2), performing a log-10 transformation on the FC values resulted in similar results to those provided in the original paper. The same 32 up-regulated genes were found.

**Table 2.**Comparison of results of study findings and analysis findings with respect to RankProd analysis for up-regulated genes

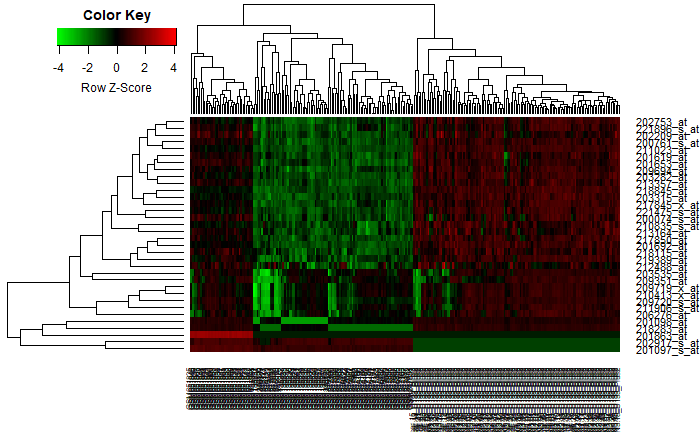
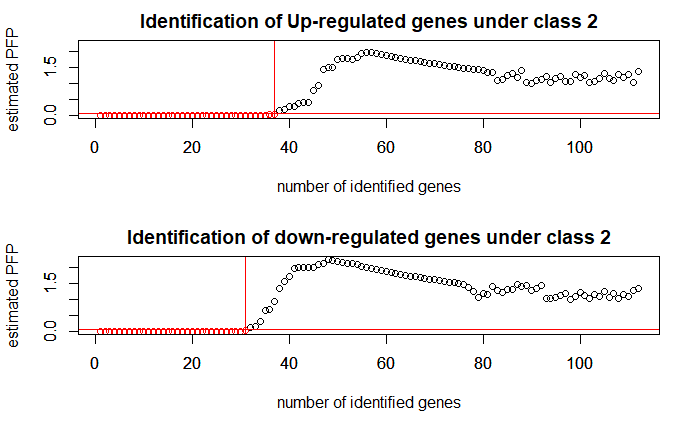
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene ID | P-value | Found FC | Found log-FC | Reported FC |
| 206731\_at | 6.6E-157 | 13.05 | 1.12 | 1.13 |
| 209243\_s\_at | 2.4E-151 | 13.42 | 1.13 | 1.13 |
| 210549\_s\_at | 8.1E-116 | 14.91 | 1.17 | 1.19 |
| 216108\_at | 1.2E-111 | 16.36 | 1.21 | 1.22 |
| 220298\_s\_at | 2.7E-82 | 18.52 | 1.27 | 1.29 |
| 212797\_at | 4.7E-82 | 20.93 | 1.32 | 1.33 |
| 207996\_s\_at | 1.0E-50 | 39.52 | 1.60 | 1.66 |
| 208762\_at | 5.7E-47 | 40.25 | 1.60 | 1.66 |
| 207257\_at | 3.3E-41 | 30.07 | 1.48 | 1.48 |
| 202035\_s\_at | 5.1E-40 | 31.81 | 1.50 | 1.50 |
| 204672\_s\_at | 3.1E-37 | 34.39 | 1.54 | 1.57 |
| 220994\_s\_at | 9.3E-36 | 35.3 | 1.55 | 1.57 |
| 213652\_at | 5.3E-34 | 41.22 | 1.62 | 1.67 |
| 213994\_s\_at | 3.3E-33 | 38.36 | 1.58 | 1.61 |
| 221606\_s\_at | 6.6E-30 | 43.53 | 1.64 | 1.68 |
| 203440\_at | 7.2E-29 | 43.14 | 1.63 | 1.65 |
| 207046\_at | 5.3E-26 | 44.5 | 1.65 | 1.66 |
| 211865\_s\_at | 2.3E-20 | 49.14 | 1.69 | 1.69 |
| 215401\_at | 2.8E-19 | 54.31 | 1.73 | 1.74 |
| 213068\_at | 4.2E-18 | 53.93 | 1.73 | 1.75 |
| 214117\_s\_at | 2.3E-14 | 60.97 | 1.79 | 1.80 |
| 34187\_at | 5.2E-13 | 63.72 | 1.80 | 1.83 |
| 213611\_at | 3.6E-11 | 72.1 | 1.86 | 1.86 |
| 209465\_x\_at | 6.2E-10 | 88.92 | 1.95 | 2.00 |
| 208606\_s\_at | 1.1E-08 | 79.91 | 1.90 | 1.92 |
| 211814\_s\_at | 2.5E-04 | 96.26 | 1.98 | 2.00 |
| 208790\_s\_at | 3.3E-04 | 101 | 2.00 | 2.00 |
| 200908\_s\_at | 5.9E-04 | 104.2 | 2.02 | 2.04 |
| 219795\_at | 6.8E-04 | 152.3 | 2.18 | 2.22 |
| 205730\_s\_at | 3.0E-03 | 108.5 | 2.04 | 2.05 |
| 202648\_at | 1.3E-02 | 118 | 2.07 | 2.10 |
| 216267\_s\_at | 2.1E-02 | 114.8 | 2.06 | 2.07 |

For the down-regulated genes, no technique proved successful in matching the FC values (Table 3). Additionally, one downregulated gene with a p-value of <10-313 was found in this analysis that was not included in the original work, resulting in 34 down-regulated genes rather than the reported 33.

**Table 3.**Comparison of results of study findings and analysis findings with respect to RankProd analysis for down-regulated genes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene ID | P-value | Found FC | Found log-FC | Reported FC |
| 202917\_s\_at | 0.0E+00 | 8061 | 3.91 | \*not in set |
| 221475\_s\_at | 1.2E-121 | 4407 | 3.64 | 0.67 |
| 203535\_at | 3.1E-110 | 5054 | 3.70 | 0.71 |
| 209719\_x\_at | 1.0E-80 | 2770 | 3.44 | 0.46 |
| 200074\_s\_at | 1.8E-74 | 2727 | 3.44 | 0.45 |
| 209351\_at | 1.2E-70 | 2346 | 3.37 | 0.39 |
| 201097\_s\_at | 1.5E-59 | 1916 | 3.28 | 0.32 |
| 217845\_x\_at | 2.2E-48 | 1623 | 3.21 | 0.25 |
| 210413\_x\_at | 2.0E-47 | 1803 | 3.26 | 0.27 |
| 210835\_s\_at | 2.0E-39 | 1359 | 3.13 | 0.17 |
| 209720\_s\_at | 5.6E-36 | 1243 | 3.09 | 0.13 |
| 200761\_s\_at | 2.7E-35 | 1248 | 3.10 | 0.13 |
| 201619\_at | 8.7E-34 | 1188 | 3.07 | 0.10 |
| 206276\_at | 3.9E-33 | 1241 | 3.09 | 0.15 |
| 201098\_at | 4.1E-29 | 1115 | 3.05 | 0.07 |
| 201653\_at | 2.8E-27 | 1089 | 3.04 | 0.07 |
| 211906\_s\_at | 5.5E-27 | 964.8 | 2.98 | 0.00 |
| 202753\_at | 5.5E-24 | 975.3 | 2.99 | 0.01 |
| 202209\_at | 3.2E-23 | 961.5 | 2.98 | 0.00 |
| 211023\_at | 3.7E-23 | 951.5 | 2.98 | 0.02 |
| 221896\_s\_at | 6.6E-20 | 905.3 | 2.96 | 954.60 |
| 213164\_at | 4.0E-15 | 731.8 | 2.86 | 787.90 |
| 201863\_at | 2.8E-14 | 721 | 2.86 | 767.20 |
| 209694\_at | 2.3E-11 | 645 | 2.81 | 688.40 |
| 218845\_at | 6.9E-10 | 601.6 | 2.78 | 675.40 |
| 203315\_at | 2.8E-08 | 547.2 | 2.74 | 584.50 |
| 213357\_at | 4.8E-08 | 563.6 | 2.75 | 612.80 |
| 217850\_at | 7.1E-08 | 550.2 | 2.74 | 592.60 |
| 218283\_at | 1.7E-07 | 537 | 2.73 | 578.60 |
| 203282\_at | 1.5E-05 | 478.2 | 2.68 | 526.70 |
| 212488\_at | 1.5E-04 | 407.8 | 2.61 | 426.30 |
| 219389\_at | 9.3E-04 | 406.9 | 2.61 | 439.20 |
| 218115\_at | 4.2E-03 | 390.8 | 2.59 | 405.00 |
| 201692\_at | 8.1E-03 | 369.2 | 2.57 | 393.00 |

### 3.2.2  Replication of analysis with created dataset

The results were not as similar when the RankProd analysis was carried out for the gene set described in the original paper using the expression values created in this work. Though the values were quantile normalized and RMA-transformed as outlined in the paper, the expression values were found to be dissimilar enough to produce different sets of both up-regulated and down-regulated genes. The up-regulated set varied more significantly, with 38 genes found, 27 of them in the original set produced in the paper. The down-regulated set was fairly consistent, finding the same 33 genes found by the analysis in section 4.1.1, which contained the 32 genes found in the original work with 1 additional. It is unclear why the up-regulated genes were found to be less replicable than the down-regulated genes. The main difference found was in the variation in pfp for these two gene sets (Fig. 1), with a larger number of genes with a lower pfp in the up-regulated set that may cause more variation in the algorithm. It is likely also due in part to a technique used in the analysis performed for the original study that was not noted in the text.

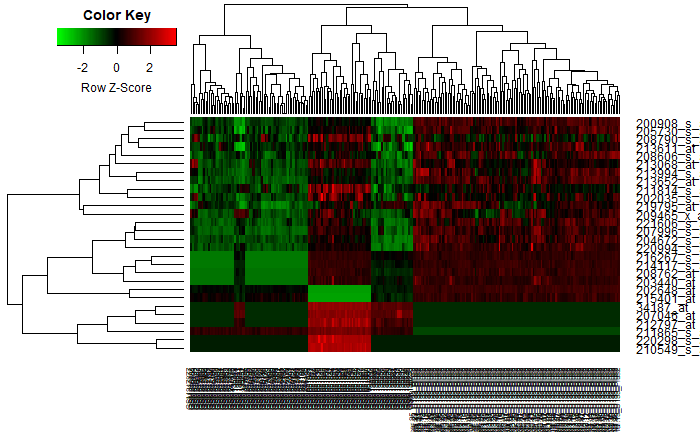
**Fig. 2: Hierarchical clustering of down-regulated genes.** Results of hierarchical clustering displayed as a heat map for the down-regulated genes. Genes show as rows, and samples as columns.

**Fig. 1: Estimated pfp for up- and down-regulated genes created from original gene set with found expression values.** The slope following the pfp cut off of 0.05 is steeper for down-regulated genes, perhaps contributing to consistency between multiple analyses.

### 3.2.3  Analysis performed with created LIMMA intersection

As the 526 genes found to be shared between the top 10,000 genes differentially expressed in each gene created too large of a set for RankProd, a curated set of 100 genes was produced by including the 3 genes that overlapped with the original study and randomly selecting 97 others. There was only 1 gene in this overlapped group that was found to be significantly up- or down-regulated in the original study, and this gene (213164\_at) was found to be significantly down-regulated in both analyses. A total of 35 up-regulated and 34 down-regulated genes were found with a p-value and a pfp of <0.05.

## Hierarchical Clustering

In the original paper, supervised and unsupervised models including Pearson Correlation Coefficient (PCC), Relief-F Feature Selection, Correlation Feature Selection, Information Gain (IG), Sequential Forward Selection, Support Vector Machines, Random Forest, and Hierarchical Clustering were used. Replication of PCC and IG was attempted, but due to time spent on multiple iterations of attempts to replicate the LIMMA and RankProd analyses, as well as the small amount of information on these methods provided in the paper, only hierarchical clustering (HC) was performed. HC utilized the average expression values obtained through this analysis, as it required normal samples not available in the appendices of the original paper. The distance function was not specified in the original work, and as such all available functions were tried and the best performance was found with it set to maximum. Quantile normalization was attempted, but led to low quality results that did not correspond with those reported in the original work, and as such the RMA normalized data was used.

**Fig. 2: Hierarchical clustering of up-regulated genes.** Results of hierarchical clustering displayed as a heat map for the up-regulated genes. Genes show as rows, and samples as columns.

Figure 2

The up-regulated and down-regulated genes were found to cluster well (Fig. 2, 3), with primary clustering appearing by gene set used and secondary clustering within each gene set. Each tree created for the up- and down-regulated genes was cut at a height of the maximum height minus 1.5, and is alluded to in the original paper. Genes in the largest remaining cluster were then considered to have been selected by the algorithm, as in the methods outlined in the original study, and correlated groups were identified. 20 highly correlated up-regulated genes were found in the original study, and 12 in this analysis. 7 overlapped between the two sets (Table 4).

**Table 4.**Comparison of results of study findings and analysis findings with respect to hierarchical clustering for up-regulated genes

|  |  |  |  |
| --- | --- | --- | --- |
| Gene ID | Identifier | Original HC | Replicated HC |
| 206731\_at | CNKSR2 | 1 | 0 |
| 209243\_s\_at | PEG3 | 1 | 0 |
| 210549\_s\_at | CCL23 | 1 | 0 |
| 216108\_at | LOC105373738 | 1 | 0 |
| 220298\_s\_at | SPATA6 | 1 | 0 |
| 212797\_at | SORT1 | 0 | 0 |
| 207996\_s\_at | LDLRAD4 | 0 | 0 |
| 208762\_at | SUMO1 | 0 | 0 |
| 207257\_at | EPO | 1 | 0 |
| 202035\_s\_at | SFRP1 | 1 | 1 |
| 204672\_s\_at | ANKRD6 | 0 | 0 |
| 220994\_s\_at | STXBP6 | 0 | 0 |
| 213652\_at | PCSK5 | 1 | 1 |
| 213994\_s\_at | SPON1 | 1 | 1 |
| 221606\_s\_at | HMGN5 | 0 | 0 |
| 203440\_at | CDH2 | 1 | 0 |
| 207046\_at | HIST2H4B | 1 | 0 |
| 211865\_s\_at | FZR1 | 0 | 0 |
| 215401\_at | AU147698 | 1 | 0 |
| 213068\_at | DPT | 1 | 1 |
| 214117\_s\_at | BTD | 1 | 0 |
| 34187\_at | AK026407 | 1 | 0 |
| 213611\_at | AQP5 | 1 | 1 |
| 209465\_x\_at | PTN | 0 | 1 |
| 208606\_s\_at | WNT4 | 0 | 1 |
| 211814\_s\_at | CCNE2 | 0 | 1 |
| 208790\_s\_at | PTRF | 0 | 1 |
| 200908\_s\_at | RPLP2 | 1 | 1 |
| 219795\_at | SLC6A14 | 0 | 1 |
| 205730\_s\_at | ABLIM3 | 1 | 1 |
| 202648\_at | TCF3 | 1 | 0 |
| 216267\_s\_at | TMEM115 | 1 | 0 |

As with the RankProd analysis, the down-regulated gene set proved more consistent. 26 highly correlated down-regulated genes were found in the original analysis. 29 were found in the replication of the analysis, with 22 of these shared with the original set (Table 5).

**Table 5.**Comparison of results of study findings and analysis findings with respect to hierarchical clustering for down-regulated genes

|  |  |  |  |
| --- | --- | --- | --- |
| Gene ID | Identifier | Original HC | Replicated HC |
| 221475\_s\_at | RPL15 | 1 | 1 |
| 203535\_at | S100A9 | 0 | 1 |
| 209719\_x\_at | SERPINB3 | 0 | 1 |
| 200074\_s\_at | GUK1 | 1 | 1 |
| 209351\_at | KRT14 | 0 | 1 |
| 201097\_s\_at | ARF4 | 1 | 0 |
| 217845\_x\_at | HIGD1A | 1 | 1 |
| 210413\_x\_at | SERPINB4 | 0 | 1 |
| 210835\_s\_at | CTBP2 | 1 | 1 |
| 209720\_s\_at | SERPINB3 | 0 | 1 |
| 200761\_s\_at | ARL6IP5 | 1 | 1 |
| 201619\_at | PRDX3 | 1 | 1 |
| 206276\_at | LY6D | 0 | 1 |
| 201098\_at | COPB2 | 1 | 0 |
| 201653\_at | CNIH1 | 1 | 1 |
| 211906\_s\_at | SERPINB4 | 0 | 1 |
| 202753\_at | PSMD6 | 1 | 1 |
| 202209\_at | LSM3 | 1 | 1 |
| 211023\_at | PDHB | 1 | 1 |
| 221896\_s\_at | HIGD1A | 1 | 1 |
| 213164\_at | SLC5A3 | 1 | 1 |
| 201863\_at | FAM32A | 1 | 0 |
| 209694\_at | PTS | 1 | 1 |
| 218845\_at | DUSP22 | 1 | 1 |
| 203315\_at | NCK2 | 1 | 1 |
| 213357\_at | GTF2H5 | 1 | 1 |
| 217850\_at | SNORD19B | 1 | 1 |
| 218283\_at | SS18L2 | 1 | 0 |
| 203282\_at | GBE1 | 1 | 1 |
| 212488\_at | COL5A1 | 1 | 1 |
| 219389\_at | SUSD4 | 1 | 1 |
| 218115\_at | ASF1B | 1 | 1 |
| 201692\_at | SIGMAR1 | 1 | 1 |

# Conclusions

This study attempted to replicate a proposed meta-analysis framework (Mei Sze Tan, 2018), looking at differential gene expression analysis of a single dataset and combined dataset, as well as hierarchical cluster on the combined dataset. While meta-analysis may be a valuable tool for identifying gene markers associated with cervical and other cancers, lack of documentation in the original study led to mixed results. A number of inconsistencies in the paper and in the associated data were found, such as reported p-value cut offs not appearing to have been used in the gene selection process and unclear normalization techniques. Further work is required to understand how such a meta-analysis process can be applied in practice, and to rectify possible errors in normalization and creation of structures for algorithms such as LIMMA and RankProd.

Acknowledgements

Parts of code framework for various algorithm implementation is owed to notes from classes with Professor Wei-Yi Cheng (Genomic Information) and Nicholas Tatonetti (Translational Bioinformatics). The forums at <https://www.biostars.org/> proved extremely helpful in understanding errors and productionizing algorithms.

References

American Cancer Society. (2018). What are the Risk Factors for Cervical Cancer? Retrieved from https://www.cancer.org/cancer/cervical-cancer/causes-risks-prevention/risk-factors.html

Bachtiary B, B. P. (2006). Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity. Clinical Cancer Research, 5632-5640.

Bolstad, B. (2018). preprocessCore: A collection of pre-processing functions. https://github.com/bmbolstad/preprocessCore

Davis, S. M. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and Bioconductor. Bioinformatics, 1846-1847.

Del Carratore, F. J. (2017). RankProd 2.0: a refactored Bioconductor package for detecting differentially expressed features in molecular profiling datasets. Bioinformatics.

Den Boon JA, P. D. (2015). Molecular transitions from papillomavirus infection to cervical precancer and cancer: role of stromal estrogen receptor signaling. Proceedings of the National Academy of Sciences of the United states of America, (pp. E3255–E3264).

E L Ivansson, I. J.-P. (2011). Pathway-based analysis of genetic susceptibility to cervical cancer in situ: HLA-DPB1 affects risk in Swedish women. Genes and Immunity, 605-614.

F. Xavier Bosch, M. M. (1995). Prevalence of Human Papillomavirus in Cervical Cancer: a Worldwide Perspective. Journal of the National Cancer Institute, 796-802.

Gautier, L. C. (2004). affy---analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 307-315.

Kejia Wu, Y. Y. (2018). Identification of key pathways and genes in the progression of cervical cancer using bioinformatics analysis. Oncology Letters, 1003-1009.

Mei Sze Tan, S.-W. C. (2018). Integrative machine learning analysis of multiple gene expression profiles in cervical cancer. PeerJ, e5285.

Miller, C. J. (2018). simpleaffy: Very simple high level analysis of Affymetrix data. Retrieved from http://bioinformatics.picr.man.ac.uk/simpleaffy/

Noordhuis MG, F. R. (2011). Involvement of the TGF-beta and beta-catenin pathways in pelvic lymph node metastasis in early-stage cervical cancer. Clinical Cancer Research, 1317-1330.

R Core Team. (2018). R: A language and environment for statistical computing. Retrieved from R Foundation for Statistical Computing: https://www.R-project.org/

Ritchie, M. P. (2015). LIMMA powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, e47.

Scotto L, N. G.-P. (2008). dentification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. Genes, Chromosomes, and Cancer, 755-765.

Warnes, G. B. (2016). gplots: Various R Programming Tools for Plotting Data. Retrieved from https://CRAN.R-project.org/package=gplots

Wong YF, C. T. (2006). Genome-wide gene expression profiling of cervical cancer in Hong Kong women by oligonucleotide microarray. International Journal of Cancer, 2461-2469.

World Health Organization. (2018). Cervical Cancer. Retrieved from Cancer: https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/