Integrative machine learning analysis of multiple gene expression profiles in cervical cancer

Mei Sze Tan 1, Siow-Wee Chang Correspond., 1, Phaik Leng Cheah 2, Hwa Jen Yap 3

1 Bioinformatics Programme, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
2 Department of Pathology, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
3 Department of Mechanical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia

Corresponding Author: Siow-Wee Chang
Email address: siowwee@um.edu.my

Although most of the cervical cancer cases are reported to be closely related to the Human Papillomavirus (HPV) infection, there is a need to study genes that stand up differentially in the final actualization of cervical cancers following HPV infection. In this study, we proposed an integrative machine learning approach to analyse multiple gene expression profiles in cervical cancer in order to identify a set of genetic markers that are associated with and may eventually aid in the diagnosis or prognosis of cervical cancers. The proposed integrative analysis is composed of three steps, namely, (i) gene expression analysis of individual dataset; (ii) meta-analysis of multiple datasets and (iii) feature selection and machine learning analysis. As a result, 21 gene expressions were identified through the integrative machine learning analysis which including seven supervised and one unsupervised methods. A functional analysis with GSEA (Gene Set Enrichment Analysis) was performed on the selected 21-gene expression set and showed significant enrichment in 9-potential-gene-expression signature, namely PEG3, SPON1, BTD and RPLP2 (upregulated genes) and PRDX3, COPB2, LSM3, SLC5A3 and AS1B (downregulated genes).
Integrative machine learning analysis of multiple gene expression profiles in cervical cancer.

Mei Sze Tan¹, Siow-Wee Chang¹*, Phaik Leng Cheah², Hwa Jen Yap³

¹Bioinformatics Programme, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia.
²Department of Pathology, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia.
³Department of Mechanical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia.

* Corresponding author

Siow-Wee Chang

siowwe@um.edu.my / changsiowwee@gmail.com

Abstract

Although most of the cervical cancer cases are reported to be closely related to the Human Papillomavirus (HPV) infection, there is a need to study genes that stand up differentially in the final actualization of cervical cancers following HPV infection. In this study, we proposed an integrative machine learning approach to analyse multiple gene expression profiles in cervical cancer in order to identify a set of genetic markers that are associated with and may eventually aid in the diagnosis or prognosis of cervical cancers. The proposed integrative analysis is composed of three steps, namely, (i) gene expression analysis of individual dataset; (ii) meta-analysis of
multiple datasets and (iii) feature selection and machine learning analysis. As a result, 21 gene expressions were identified through the integrative machine learning analysis which including seven supervised and one unsupervised methods. A functional analysis with GSEA (Gene Set Enrichment Analysis) was performed on the selected 21-gene expression set and showed significant enrichment in 9-potential-gene-expression signature, namely *PEG3, SPON1, BTD* and *RPLP2* (upregulated genes) and *PRDX3, COPB2, LSM3, SLC5A3* and *AS1B* (downregulated genes).

**Introduction**

Cervical cancer, ranked 4\textsuperscript{th} most common cancer [1] among women worldwide, caused 279,311 mortality cases in the year 2015 [2]. It is the 3\textsuperscript{rd} most common cancer (7.7\%) among Malaysian females [3]. Most of the development of cervical cancer is linked to the infection of the Human Papillomavirus (HPV) that passes to the human host during sexual activity. The other mentioned risk factors include smoking, imbalanced diet and family history [4-6]. However, as in many cancers, it is conceivable that genetic pathways play a role in the development of cervical cancer.

Microarray technology has been used to simultaneously identify a large number of gene expressions that may be associated with a particular cancer. Reports of prognostically predictive genes have been identified by whole genome microarray in cancer. Hu et al. [7] showed that miR-200a and miR-9 were associated with cervical cancer outcome while Aziz et al. [8], reported on 19 common genes that were predictive in colorectal cancers. Meta-analysis is touted to solve the problems of low statistical power of studies, to be able to analyze intersected genes between datasets, extract the optimum values across multiple datasets and identify the most commonly
expressed genes [9] [10]. Based on our literature review, we believed that this is one of the first few studies which involves meta-analysis using multiple cervical cancer microarray datasets.

Despite many approaches were introduced in the previous meta-analysis study, biased issue that is caused by merging datasets from different platforms and experiments is a hot topic among the field. The studies in [11-14] carried out meta-analysis in various type of cancers to find the significant genes in certain diseases such as breast cancer, paediatric B-acute lymphoblastic leukaemia (B-ALL), ovarian cancer and etc. Trendily, most of these studies were done by using the original datasets that varied from different platforms. As the original datasets contained noisy data such as the extreme values, thus the bias effects from each of the respective datasets might be strong.

Furthermore, until to-date, most of the previous studies are either using statistical analysis methods [11-14] or machine learning-based methods [12-19] in carrying out the meta-analysis. Due to different measurements being used in analysing the expression activities in the genes, there is no way in comparing the feasibility of both methods in a meta-analysis. Hence, by combining both the statistical and machine learning methods, meaningful insights could be discovered. Favourable features from both of the methods are integrated and fully utilized in carrying out the meta-analysis in this study.

To address the shortcomings of meta-analysis we attempted to build a system which integrates machine learning with meta-analysis to improve the quality of the data acquired. The whole process encompasses (i) Gene expression analysis of individual dataset; (ii) meta-analysis of multiple datasets; and (iii) feature selection (FS) & machine learning analysis. We report our findings using our proposed integrative analysis in identifying genes which could be integral in
the development of cervical cancers, a cancer in which understanding has moved significantly forward and awaits unravelling of possible genetic intermediaries in the carcinogenetic process.

Materials and Methods

In brief, our approach commences with the gene expression analysis of individual cervical cancer microarray datasets obtained from publicly available sources and to detect the genes that perform most significantly in each of the dataset. The most significant gene sets from each dataset are combined and intersected to form a larger dataset in the first stage. Microarray gene expression meta-analysis is performed using the combined dataset in the second stage. For the final stage, several FS and machine learning methods as well as a functional analysis by gene set enrichment analysis (GSEA) will be carried out to identify a more precise set of potential gene markers in cervical cancer biology.

Microarray Datasets

All the data used are from the public website National Center Biotechnology Information Gene Expression Omnibus (NCBI GEO) (http://www.ncbi.nlm.nih.gov/geo). A search using the keywords “cervical cancer” and “microarray” was carried out. To ensure the accurateness and purity of the resultant 104 (before 5th January 2017) datasets, filtration was performed by manually reviewing each dataset using the following criteria. To gain entry into this study, the dataset must 1) contain expression profiling by array types, 2) use CEL raw files, and 3) be studies that involve cervical cancer as the main subject. Finally, only four cervical cancer microarray gene expression datasets (see Table 1) could be entered for further analysis.

Table 1 Details of the four microarray datasets used in this research.
All analyses of the data were performed in R programming environment (version 3.3.2).

**Integrative approach of meta-analysis and machine learning**

The aim of the proposed integrative approach is to fully utilize the benefits from both of the statistical analysis and the machine learning approaches in carrying out gene expression meta-analysis, in order to find out the most significant gene markers in cervical cancer. Generally, the proposed integrative analysis is composed of three stages, namely, (i) gene expression analysis of individual dataset; (ii) meta-analysis of multiple datasets and (iii) feature selection and machine learning analysis. The proposed integrative method is illustrated as in Figure 1. At the first stage, gene expression analysis of individual dataset was carried out where 10,000 differentially expressed genes from each dataset were selected by using Linear Models for Microarray Analysis (LIMMA). This process was repeated on four datasets. In stage (ii), meta-analysis of four datasets was carried out. Intersection process was done in order to obtain the intersected genes among the four datasets. A set of upregulated and downregulated genes were selected and ranked accordingly by using Rank Prod at the end of stage (ii). In stage (iii), eight FS and machine learning approaches were applied in order to identify the common optimum genes.

**Figure 1 Framework for the proposed integrative approach of meta-analysis and machine learning in gene expression profiling of cervical cancer.**

*DE- differential expression value

Stages of the integrative approach are as detailed below:

**Stage (i) Gene expression analysis of individual dataset**

The data were first filtered to select out cancer samples (removal of normal samples) to ensure that the activities of the genes presented in the normal samples would not affect the analysis. Among
the four sets of microarray data used in our study, only microarray data of GSE9750 [22] is based on the Affymetrix Human Genome U133A Array [HG-U133A] platform. The other three datasets are based on Affymetrix Human Genome U133 Plus 2.0 Array [HG-U133_Plus_2]. Therefore, it is necessary to pre-process and normalize all of the raw level probe data across the platforms to reduce bias caused by the different platforms. Quantile algorithm were applied to normalize the data and the data were log2-transformed according to all the samples median.

Then, gene expression analysis was performed on each of the microarray datasets. RMA normalization from \textit{affy} function in Bioconductor package R was carried out for every sample in the dataset to create expression values that are normalized and background corrected.

After that, differential gene expression analysis was carried out to select the top 10000 differentially expressed genes from individual dataset using Linear Models for Microarray Analysis (LIMMA) [24]. In this stage, the rare and over-extremely expressed genes in every dataset were eliminated by using moderated \textit{t}-statistics of LIMMA so that bias errors from each dataset could be reduced when combining the datasets to become a larger dataset.

\textit{R} package of \textit{limma} was used to rank and select out the top differential expressed genes. The differentially-expressed genes were filtered using adjusted \textit{p}-value of <0.05 and ranked based according to the \textit{p}-values. In this study, the same gene expression analysis was repeated on all the four datasets.

Hence, this stage is treated as quality control of the dataset which is going to use in the stage (ii) analysis as genes produced by this stage are consistently different between the group according to the LIMMA statistics.

\textbf{Stage (ii): Meta-analysis for multiple datasets}
The aim of this stage is to select the genes that show upregulated and downregulated activities in cervical cancer. Four microarray gene expression datasets that contained top 10,000 genes each (selected in the stage(i)) were combined in the intersection process by identifying the intersected genes found among the four datasets and merging all samples to form a single, larger sample size dataset (meta-dataset).

Next, meta-analysis was conducted by using RankProd [25] package in R to analyse the intersected genes. With the default value of \( p < 0.05, \text{FDR}<0.05 \) as the cut-off point, RankProd ranked the genes into two group: upregulated (FC>1, \( p\text{h}<0.01 \)) and downregulated genes (FC<1, \( p\text{h}<0.01 \)).

**Stage (iii) Feature selection (FS) and machine learning methods**

Stage (iii) is to select the optimum genes to ensure the final set of genes selected is with higher standard and high quality in term of genetic evaluation. This is also to fully utilize the measurements of different machine learning methods (supervised and unsupervised) by collaborating them in selecting the significant genes. The machine learning methods chosen in this study are based on the literature review done which obtained good result in the related studies [26-33].

Seven supervised machine learning FS methods and an unsupervised were used in this study. Among the seven supervised machine learning FS methods, four filter approaches which are Pearson Correlation Coefficient(PCC), Relief-F Feature Selection (Relief-F), Correlation Feature Selection (CFS) and Information Gain (IG) are used in this study for their simplicity and high in good succession in performing FS. The features (genes) that selected by the filter approach are based on the relevance of the genes with the surrounding data [34,35]. Besides that, one wrapper approach, Sequential Forward Selection (SFS) was carried out in this study as it acts as a searching
algorithm that could find a subset of genes that is maximum according to the objective function (classification performance) [34,35]. Furthermore, this study also applied two embedded approaches which are Support Vector Machine-based Recursive Feature Elimination (SVM-RFE) and Random Forest (RF) which embed variable selection as part of their learning process and evaluate them according to the classification result of specific machine learning method [34,35]. Hierarchical Clustering (HC) was applied as the unsupervised model in this study for its ability to select the genes without the knowledge of the class labels [34].

As the supervised machine learning methods involves the classification according to the class labels, the normal samples were added back at this stage (ii) so that the “learning procedures” of the machine learning can be completed.

The machine learning approaches applied in this study are more on exploratory models which the performances of the machine learning methods are not being emphasize here. The main objective of this stage is to obtain the optimum genes that commonly selected using different machine learning methods, regardless of the differences in the algorithm used. The results (selected genes) of different FS and machine learning will be combined and converted in the scoring matrix. Scores will be determined and counted by the frequency of the genes that was being selected by various FS and machine learning methods.

**Functional Pathway Analysis**

**Gene Set Enrichment Analysis (GSEA)**

GSEA ([http://software.broadinstitute.org/gsea/index.jsp](http://software.broadinstitute.org/gsea/index.jsp)) [36,37] computationally evaluates the statistical significance of a priori defined set of genes with their biological states (i.e, phenotypes). All the genes in the datasets are ranked according to their differential expression between
phenotype groups. In this study, normal tissue samples are also added into the dataset to enable
the GSEA to calculate the Enrichment Scores (ES) between the phenotypes. A computational
annotated gene set (C4) that consists of large collections of cancer-orientated microarray data is
provided by Molecular Signature Database (MSigDB) and linked to GSEA in order to calculate
the ES on our selected set of genes. ES is computed for every set of genes by going through the
rank list from the top. A gene set will have a higher ES if the encountered gene in the gene input
list is a member of the annotated gene set. A gene set is said to be significant if the normalized
enrichment scores (NES) is high and its corresponding false discovery rate (FDR) is low. In this
study, FDR value of <0.25 was set as the threshold to consider the significance of the gene
enriched. Furthermore, pathway network could be visualized through Cytoscape [38] software
based on the GSEA results.

Result

Stage (i): Gene Expression Analysis of Individual Dataset

LIMMA was used to perform the analysis in order to select out the top 10,000 differentially
expressed genes from each of the individual dataset (Results showed as in Appendix 1).

LIMMA tests the average difference in log expression levels of two groups per gene by using
moderated t-test. Therefore, genes selected by LIMMA are moderately different with average
differential values, rather than extremely different.

Stage (ii): Meta-Analysis of Multiple Dataset

In the intersection process, 117 intersected genes among four cervical cancer datasets were
identified. All cervical cancer samples from four datasets (total of 142 samples) were combined to
form a meta-dataset with these 117 intersected genes. The meta-dataset is attached as in Appendix 2.

Next, the analysis of the 117 genes by \textit{RankProd} identified 32 upregulated genes (FC > 1, pfp < 0.01) and 33 down regulated genes (FC < 1, pfp < 0.01), with p-value<0.05 and FDR of 0.05 as the cutoff points.

\textit{RankProd} estimated the FDR for each gene and ranked the gene according to the value that has equal or smaller p-value in the whole list of genes (estimated percentage of false predictions). As the product of \textit{RankProd} is affected more by extreme expression values, genes that are extremely expressed in at least one sample in the dataset tends to be selected. Table 2 and Table 3 show the results of \textit{RankProd} analysis.

**Table 2 Results of \textit{RankProd} analysis- 32 Upregulated genes.**

**Table 3 Results of \textit{RankProd} analysis- 33 Downregulated genes.**

*Notes: Description of Table 2 and Table 3: gene.index- index of genes; RP/Rsum- computed Rank Product statistics; FC: (class1/class2)- average expression levels’ computed fold change under two conditions (upregulated class and downregulated class); pfp- estimated false positive predictions value of the genes; and P-value- estimated p-value of each gene.

**Stage (iii): Feature Selection and Machine Learning Methods**

To find the optimum set of genes, eight FS and machine learning methods namely PCC, Relief-F, CFS, IG, SFS, SVM-RFE, RF and HC, were carried out using the set of genes resulting from the stage (ii) of meta-analysis.

For HC analysis, the expression patterns of the 32 upregulated genes are shown in the heatmap and were further sub-clustered by cutting tree at the maximum heights of 1.5 of the row cluster. As a result, a cluster group that contained 20 genes was found to be most correlated to each other.
The same rule was applied to the 33 downregulated genes. The most correlated group that contained 26 genes was also identified (Figure 2 (b)). The genes that were selected in the cluster that had the highest scores are presented in Table 4 and Table 5 respectively.

Figure 2 Hierarchical clustering of the selected genes. (a) Hierarchical clustering of the 32 upregulated genes. (b) Hierarchical clustering of the 33 downregulated genes. In the figures, each gene is represented by the rows and the samples are represented by the columns. The dendrogram at the side indicates the relation between the pattern of the gene expression while the top-dendrogram indicates the relation between the samples used. The level of expression of the genes, relative to the mean of the gene across all samples, is indicated by the colour key, with the green representing the higher expression of the genes. The colour bar on the left side represents the cluster of the genes after they have been cut at 1.5 of the height of the tree so that the clustering of the genes is seen more clearly. The colour of the bar is indicated by the colour key similarly.

As for the scoring measures, the final score obtained by each of the genes is the frequency it was selected by the listed FS and machine learning methods (1 score for selected, 0 score for not selected). The higher the score of a gene, the more frequently the gene was selected by the different FS methods, and the higher the potential for the gene to be a gene marker of cervical cancer. Table 4 and Table 5 show the set of genes selected by the FS methods. The up- and downregulated genes with score ≥ 6 (full score = 8, threshold = 75%) were identified and selected. In total, 21 gene expression set consisted of 9 upregulated genes and 12 downregulated genes were selected.

Table 4 Ranking of each gene (upregulated) using proposed FS methods.
*The 9 upregulated genes selected by most (≥ 6) of the methods are CNKSR2, PEG3, SPATA6, SPON1, BTD, AK026407, RPLP2, TCF3 and TMEM115.

Table 5 Ranking of each gene (downregulated) using proposed FS methods.
*The 12 downregulated genes selected by most (≥ 6) of the methods are RPL15, HIGD1A, ARL6IP5, PRDX3, COPB2, CNIH1, PSMD6, LSM3, PDHB, SLC5A3, PTS and ASF1B. As gene HIGD1A is selected twice with different ID REF, it only considered once here.

Functional Pathway analysis with GSEA
GSEA was carried out on the selected 21-gene expression set for functional analysis to measure the statistical significance of their biological phenotypes. In our study, we are assessing the enrichment of the selected 9-upregulated and 12-downregulated genes on their natural phenotypes, which are tumour and non-tumour. Therefore, samples of normal cells that were filtered at the beginning of the study were added back to complete the study. A computational gene set (C4) consisted of large collections of cancer-orientated microarray data was downloaded from MSigDB to calculate the enrichment scores of the 9-upregulated genes and 12-downregulated genes among the annotated gene sets of C4. We evaluated the normalized enrichment scores (NES) and the corresponding false discovery rate (FDR) of each gene on all of the samples used in this study. In this study, a gene will be rejected when the FDR q-value is more than 0.25 (not significant). As NES values are dependent on the permutation method and size of the expression dataset, there is no standard threshold for this value in analysing the gene. A gene could match multiple sets of gene sets in the database, hence we selected the highest NES values for each respective gene as shown in Table 6 and Table 7. The details of GSEA report on the annotation of the 21-gene expression set with the description of each gene sets are attached in Appendix 3.

Table 6 Results of GSEA analysis using the 9 upregulated-genes.
*There was no matching of gene AK026407, TCF3 and TMEM115 with the MSigDB.

Table 7 Results of GSEA analysis using the 12 downregulated-genes.
*There was no matching of gene ARL6IP5, CNIH1 and PTS with the MSigDB.

Among the 9 upregulated genes, 6 genes were matched with the cancerous gene set collections, as shown in Table 6. Notably, PEG3, SPON1, BTD and RPLP2 passed the defined threshold of FDR < 0.25, indicating the potential of these genes in showing their significances in cervical cancer. As there was very limited information regarding genes AK026407, TCF3 and TMEM115, there are no
matching genes with the MSigDB. Besides that, 9 downregulated genes were matched with the C4
gene sets in MSigDB (Table 7). This showed that the 9 genes were enriched and highly related to
cancers. Among enriched genes, five downregulated genes, PRDX3, COPB2, LSM3, SLC5A3 and
AS1B, passed the default threshold with $FDR < 0.25$. These also provided some attraction points
for the selected downregulated genes in proving their significances in cervical cancer. However,
there are no matching of the genes ARL6IP5, CNIH1 and PTS with the MSigDB due to the low
information found regarding these genes.

Figure 3 and Figure 4 show the pathway network constructed according to the GSEA report
generated in Appendix 3. The threshold that used in the pathway analysis was set to $FDR<0.25$ in
order to standardize with GSEA. Each of the nodes in the networks represented the gene sets that
the genes involved in. The genes that showed enrichment with the gene sets were indicated in the
colour nodes while the line connecting the nodes are the pathways that the enriched genes found
in.

**Figure 3** Pathway network of the upregulated genes.

**Figure 4** Pathway network of the downregulated genes.

In Figure 3, the upregulated genes PEG3, SPON1, BTD and RPLP2 were shown to be related with
the pathway of gene expression, disease or cancer, and protein metabolism. Similarly, as shown in
Figure 4, downregulated genes PRDX3, COPB2, LSM3, SLC5A3 and ASF1B were involving in
the gene expression, disease or cancer, protein metabolism and cell cycle pathways.

**Discussion**
Integrative approach of meta-analysis and machine learning

The aim of this study was to identify a set of more precise, accurate and reliable potential cervical cancer gene markers. Generally, there are three stages of analysis in this study: (i) gene expression analysis of individual dataset; (ii) meta-analysis of multiple datasets and (iii) FS and machine learning analysis.

At the first stage, individual analysis of each cervical cancer microarray gene expression dataset was carried out using LIMMA package in R. In LIMMA, the average difference of log expression levels between two groups of information provided for each gene was calculated using moderated t-statistic, to ensure that the final list of genes produced consists of genes that are consistently different between the groups. In other words, LIMMA avoids choosing the extremely different genes in the datasets. In this study, Benjamini and Hochberg’s method in controlling the FDR value was used as the default adjustment method of the adjusted p-value. The adjusted p-value is set to be <0.05 in expecting the proportion of the false discoveries in the selected gene group to be less than 5%.

In the meta-analysis stage, RankProd package was chosen for use for two reasons; to analyse the combination of datasets from different origins and select genes with higher biological significances. RankProd package is able to avoid errors that are caused from batch and platform differences in combining expression values of various studies, by performing ranking of the genes by their differential expression [39]. RankProd estimated the gene ranking by randomly permutating the list of genes. False discovery rates (FDR) for the whole list of genes which are equal or smaller than the p-value (known as false positive prediction, pfp) were calculated by this package and the genes were ranked accordingly.
The selection of genes in the RankProd analysis in the stage (ii) was based on the criteria of FC>1 (upregulated genes) or FC<1 (downregulated genes). In log expression level, a gene is said to be significantly upregulated when the FC expression is >1 while it is significantly downregulated when the FC <1. An FC with value of two simply brings the meaning that the particular genes expressed twice more than the genes in the original control. Besides that, p-value is set to <0.01 in order to select the genes that sited on the top 1% of the truly significant genes among all the genes to be studied.

Following meta-analysis, 32 most significantly up-regulated and 33 down-regulated genes were selected.

To further narrow down the range of the genes selected and also to fully adopt the benefits from the statistical and also the machine-based analysis method, four filter approaches (PCC, Relief-F, CFS and IG), a wrapper approach (SFS), two embedded approach (SVM-RFE and RF) and an unsupervised approach of HC were carried out. As the algorithms in selecting genes by each method were different, the resultant genes selected by each method varies. In this study, most of the parameters set are based on the default values or the minimum values required as the focus is on the most significant genes that selected by the machine learning method and not the predictive performances of the machine learning methods).

We ensured the accuracy of our results by creating a scoring system to calculate the score of each gene. The threshold for the scoring system was set to 75%, where only the genes with score ≥ 6 (out of 8) would be selected. In total, 21-gene expression set which consisted of 9 upregulated genes and 12 downregulated genes were selected. The upregulated genes that were most commonly selected by the stated FS methods were PEG3, SPON1, BTD, RPLP2, CNKSR2, SPATA6, AK026407, TCF3 and TMEM115 while the FS-commonly-selected-downregulated-
genes were PRDX3, COPB2, LSM3, SLC5A3, ASF1B, RPL15, HIGD1A, ARL6IP5, CNIH1, PSMD6, PDHB, and PTS.

**Functional Pathway Analysis - GSEA**

According to the annotation of the 21 genes (9 upregulated genes and 12 downregulated genes) to the C4 cancerous gene set collection in MSigDB GSEA, it showed that six upregulated genes (CNKSR2, PEG3, SPATA6, SPON1, BTD and TCF3) and nine downregulated genes (RPL15, HIGD1A, PRDX3, COPB2, PSMD6, LSM3, PDHB, SLC5A3 and ASF1B) were enriched with the computational cancerous gene set. Particularly, 4 upregulated genes (SPON1, PEG3, RPLP2 and BTD) and 5 downregulate genes (PRDX3, COPB2, LSM3, SLCA3 and ASF1B) showed significant enrichment to the cancerous modules and cancerous genes neighbourhood gene sets with the FDR<0.25. This indicates that these 9 genes were highly correlated with the formation of cancer and were selected as the 9-potential-gene-expression signature in this study.

The pathway network (Figure 3 and Figure 4) were constructed using the GSEA report that attached as in Appendix 3. The networks shown were based on the correlation between the gene sets. Similar with GSEA, the threshold was set to FDR<0.25 in this analysis. By comparing the map constructed and the GSEA result, it can be clearly seen that both results are supporting each other. The genes that showed significant enrichment with the C4 gene sets were shown and constructed in the network too. Moreover, the genes involved in the pathways of gene expression, disease or cancer, protein metabolism and cell cycle (Figure 3 and Figure 4). This further strengthened the role of the selected genes to be the 9-potential signature in relation with cervical cancer development.
The 9-potential signature will be discussed and supported by the previous studies done in the following section.

Upregulated genes are the genes that are known to be overexpressed during the gene regulation stage, causing the development of the tumour or potential cancerous cell. The most significant upregulated genes set that formed after the GSEA analysis are \textit{PEG3}, \textit{SPON1}, \textit{BTD} and \textit{RPLP2}. Table 8 shows the function of the upregulated genes selected, together with the related diseases and description.

\textbf{Table 8 Function, related diseases and description of the most significant genes (Upregulated).}

Despite the focus of gene expression research in cancers are towards the upregulated genes, the role of downregulated genes should not be ignored. Lu et al. [56] found that in general, tumours contained more downregulated miRNAs compared with the normal tissues, leading them to hypothesize that down-regulated expression of some miRNAs may play a key role in the development of tumours and that gene downregulation is as important an aspect to be studied as up-regulation.

The most significant identified downregulated genes in cervical cancer in this study are \textit{PRDX3}, \textit{COPB2}, \textit{LSM3}, \textit{SLC5A3} and \textit{ASF1B}. The biological functions, related diseases and description are shown in Table 9.

\textbf{Table 9 Function, related diseases and description of the most significant genes (Downregulated).}

Our integrated approach of meta-analysis and machine-learning method successfully identified the potential genetic markers in cervical cancer, along with the evidence provided from the GSEA results and the previous study done. The identified genes can be considered as potential gene
signatures in cervical cancer. However, further evaluation using other experimental methods are needed in order to validate the significance of the 9-potential-gene expression cervical cancer signatures.

**Conclusion**

In this study, we proposed an integrative machine learning analysis of multiple gene expression profiles in cervical cancer. This integrative analysis is divided into three stages which are (i) gene expression analysis of individual dataset; (ii) meta-analysis of multiple datasets and (iii) FS & machine learning analysis with PCC, Relief-F, CFS, IG, SFS, SVM-RFE, RF and HC. As a result, 9 upregulated genes and 12 downregulated genes, were selected through the three-step analysis. The result of the functional pathway analysis in GSEA showed the significant enrichment in 9-potential-gene-expression signature namely *PEG3, SPON1, BTD, RPLP2 (as upregulated genes)* and *PRDX3, COPB2, LSM3, SLC5A3, ASF1B (as downregulated genes)*. More experiments are needed to confirm the significance of the selected 9-potential-gene-expression signature in cervical cancer.

**Reference**

1. Cervical cancer statistics. World Cancer Research Fund International. *Available at http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/cervical-cancer-statistics* (accessed 3rd January 2017).
2. World Health Organization. (2016). *World Health Statistics 2016: Monitoring Health for the SDGs Sustainable Development Goals*. World Health Organization.
3. AbM A, Nor Saleh I, and Noor Hashima A. 2016. Malaysian National Cancer Registry Report 2007–2011. Malaysia Cancer Statistics, Data and Figure. National Cancer Institute, Ministry of Health, Putrajaya, Malaysia.

4. Plummer M, Herrero R, Franceschi S, Meijer CJ, Snijders P, Bosch FX, de Sanjosé S, and Muñoz N. 2003. Smoking and cervical cancer: pooled analysis of the IARC multi-centric case–control study. Cancer Causes and Control 14:805-814.

5. Ames BN, and Gold LS. 1997. The causes and prevention of cancer: gaining perspective. Environmental Health Perspectives 105:865.

6. Lee SA, Kang D, Seo SS, Jeong JK, Yoo KY, Jeon YT, Kim JW, Park NH, Kang SB, and Lee HP. 2003. Multiple HPV infection in cervical cancer screened by HPVDNAChip™. Cancer letters 198:187-192.

7. Hu X, Schwarz JK, Lewis JS, Huettner PC, Rader JS, Deasy JO, Grigsby PW, and Wang X. 2010. A microRNA expression signature for cervical cancer prognosis. Cancer research 70:1441-1448.

8. Aziz NAA, Mokhtar NM, Harun R, Mollah MMH, Rose IM, Sagap I, Tamil AM, Ngah WZW, and Jamal R. 2016. A 19-Gene expression signature as a predictor of survival in colorectal cancer. BMC medical genomics 9:58.

9. Warnat P, Eils R, and Brors B. 2005. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. BMC bioinformatics 6:265.

10. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, and Causton HC. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature genetics 29:365-371.
11. Alles MC, Gardiner-Garden M, Nott DJ, Wang Y, Foekens JA, Sutherland RL, Musgrove EA, and Ormandy CJ. 2009. Meta-analysis and gene set enrichment relative to er status reveal elevated activity of MYC and E2F in the “basal” breast cancer subgroup. PloS one 4:e4710.

12. Chow, Y. P., Alias, H., & Jamal, R. (2017). Meta-analysis of gene expression in relapsed childhood B-acute lymphoblastic leukemia. BMC cancer, 17(1), 120.

13. Dong, H., Hong, S., Xu, X., Xiao, Y., Jin, L., & Xiong, M. (2010, May). Meta-analysis and Network Analysis of Five Ovarian Cancer Gene Expression Dataset. In Computational Science and Optimization (CSO), 2010 Third International Joint Conference on (Vol. 1, pp. 242-246). IEEE.

14. Li, W. X., Dai, S. X., Wang, Q., Guo, Y. C., Hong, Y., Zheng, J. J., ... & Huang, J. F. (2016). Integrated analysis of ischemic stroke datasets revealed sex and age difference in anti-stroke targets. PeerJ, 4, e2470.

15. Grützmann R, Boriss H, Ammerpohl O, Lüttges J, Kalthoff H, Schackert HK, Klöppel G, Saeger HD, and Pilarsky C. 2005. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. Oncogene 24:5079-5088.

16. Liu, B., Cui, Q., Jiang, T., & Ma, S. (2004). A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC bioinformatics, 5(1), 136.

17. Obayashi, T., & Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA research, 16(5), 249-260.

18. Wang, Y., & Makedon, F. (2004, August). Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data.
19. Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1), 3.

20. Noordhuis MG, Fehrmann RS, Wisman GBA, Nijhuis ER, van Zanden JJ, Moerland PD, van Themaat EVL, Volders HH, Kok M, and Klaske A. 2011. Involvement of the TGF-β and β-catenin pathways in pelvic lymph node metastasis in early-stage cervical cancer. *Clinical Cancer Research*.

21. Bachtiary B, Boutros PC, Pintilie M, Shi W, Bastianutto C, Li J-H, Schwock J, Zhang W, Penn LZ, and Jurisica I. 2006. Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity. *Clinical Cancer Research* 12:5632-5640.

22. Scotto L, Narayan G, Nandula SV, Arias-Pulido H, Subramaniyam S, Schneider A, Kaufmann AM, Wright JD, Pothuri B, and Mansukhani M. 2008. Identification 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* 47:755-765.

23. Den Boon JA, Pyeon D, Wang SS, Horswill M, Schiffman M, Sherman M, Zuna RE, Wang Z, Hewitt SM, and Pearson R. 2015. Molecular transitions from papillomavirus infection to cervical precancer and cancer: Role of stromal estrogen receptor signaling. *Proceedings of the National Academy of Sciences* 112:E3255-E3264.

24. Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* 3:1-25.
25. Breitling R, Armengaud P, Amtmann A, and Herzyk P. 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS letters 573:83-92.

26. Urbanowicz RJ, Meeker M, LaCava W, Olson RS, and Moore JH. 2017. Relief-based feature selection: introduction and review. arXiv preprint arXiv:171108421.

27. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, and Mewes HW. 2005. Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry 29:37-46.

28. Martín-Valdivia MT, Díaz-Galiano MC, Montejo-Raez A, and Urena-Lopez L. 2008. Using information gain to improve multi-modal information retrieval systems. Information Processing & Management 44:1146-1158.

29. Guyon I, Weston J, Barnhill S, and Vapnik V. 2002. Gene selection for cancer classification using support vector machines. Machine learning 46:389-422.

30. Breiman L. 2001. Random forests. Machine learning 45:5-32.

31. Díaz-Uriarte R, and De Andres SA. 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics 7:3.

32. Sneath PH. 1973. The principles and practice of numerical classification. Numerical taxonomy 573.

33. Eisen MB, Spellman PT, Brown PO, and Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95:14863-14868.

34. Chandrashekar G, and Sahin F. 2014. A survey on feature selection methods. Computers & Electrical Engineering 40:16-28.
35. Saeys Y, Inza I, and Larrañaga P. 2007. A review of feature selection techniques in bioinformatics. *bioinformatics* 23:2507-2517.

36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, and Lander ES. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102:15545-15550.

37. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, and Laurila E. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* 34:267-273.

38. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13:2498-2504.

39. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, and Chory J. 2006. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 22:2825-2827.

40. Nye MD, Hoyo C, Huang Z, Vidal AC, Wang F, Overcash F, Smith JS, Vasquez B, Hernandez B, and Swai B. 2013. Associations between methylation of paternally expressed gene 3 (PEG3), cervical intraepithelial neoplasia and invasive cervical cancer. *PLoS one* 8:e56325.

41. Yamaguchi A, Taniguchi M, Hori O, Ogawa S, Tojo N, Matsuoka N, Miyake S-i, Kasai K, Sugimoto H, and Tamatani M. 2002. Peg3/Pw1 is involved in p53-mediated cell death pathway in brain ischemia/hypoxia. *Journal of Biological Chemistry* 277:623-629.
42. Jiang X, Yu Y, Yang HW, Agar NY, Frado L, and Johnson MD. 2010. The imprinted gene PEG3 inhibits Wnt signaling and regulates glioma growth. Journal of Biological Chemistry 285:8472-8480.

43. Burstyn-Cohen T, Tzarfaty V, Frumkin A, Feinstein Y, Stoeckli E, and Klar A. 1999. F-spondin is required for accurate pathfinding of commissural axons at the floor plate. Neuron 23:233-246.

44. Pagnotta SM, Laudanna C, Pancione M, Sabatino L, Votino C, Remo A, Cerulo L, Zoppoli P, Manfrin E, and Colantuoni V. 2013. Ensemble of gene signatures identifies novel biomarkers in colorectal Cancer activated through PPARγ and TNFα signaling. PloS one 8:e72638.

45. Zhai Y, Bommer GT, Feng Y, Wiese AB, Fearon ER, and Cho KR. 2010. Loss of estrogen receptor 1 enhances cervical cancer invasion. The American journal of pathology 177:884-895.

46. Faith M, and Abraham P. 2013. An update on diagnostic value of biotinidase: From liver damage to cancer: Minireview. Biomedical Research 24.

47. Huang L, Zheng M, Zhou QM, Zhang MY, Jia WH, Yun JP, and Wang HY. 2011. Identification of a gene-expression signature for predicting lymph node metastasis in patients with early stage cervical carcinoma. Cancer 117:3363-3373.

48. So AK-C, Kaur J, Kak I, Assi J, MacMillan C, Ralhan R, and Walfish PG. 2012. Biotinidase is a novel marker for papillary thyroid cancer aggressiveness. PloS one 7:e40956.
49. Scheerger, and Zempleni. 2003. Expression of oncogenes depends on biotin in human small cell lung cancer cells NCI-H69. International journal for vitamin and nutrition research 73:461-467.

50. Kang U-B, Ahn Y, Lee JW, Kim Y-H, Kim J, Yu M-H, Noh D-Y, and Lee C. 2010. Differential profiling of breast cancer plasma proteome by isotope-coded affinity tagging method reveals biotinidase as a breast cancer biomarker. BMC cancer 10:114.

51. Zhang Y-Z, Zhang L-H, Gao Y, Li C-H, Jia S-Q, Liu N, Cheng F, Niu D-Y, Cho WC, and Ji J-F. 2011. Discovery and validation of prognostic markers in gastric cancer by genome-wide expression profiling. World journal of gastroenterology: WJG 17:1710.

52. Sharp M, Adams S, Elvin P, Walker R, Brammar W, and Varley J. 1990. A sequence previously identified as metastasis-related encodes an acidic ribosomal phosphoprotein, P2. British journal of cancer 61:83-88.

53. Mao-De L, and Jing X. 2007. Ribosomal proteins and colorectal cancer. Current genomics 8:43-49.

54. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, and Ferrando AA. 2005. MicroRNA expression profiles classify human cancers. nature 435:834-838.

55. Safaeian M, Hildesheim A, Gonzalez P, Yu K, Porras C, Li Q, Rodriguez AC, Sherman ME, Schiffman M, and Wacholder S. 2012. Single nucleotide polymorphisms in the PRDX3 and RPS19 and risk of HPV persistence and cervical precancer/cancer. PloS one 7:e33619.
56. Jiang Y, Sang Y, and Qiu Q. 2017. microRNA-383 mediates high glucose-induced oxidative stress and apoptosis in retinal pigment epithelial cells by repressing peroxiredoxin 3. American journal of translational research 9:2374.

57. Kim K, Yu M, Han S, Oh I, Choi Y-J, Kim S, Yoon K, Jung M, and Choe W. 2009. Expression of human peroxiredoxin isoforms in response to cervical carcinogenesis. Oncology reports 21:1391-1396.

58. Hu J-X, Gao Q, and Li L. 2013. Peroxiredoxin 3 is a novel marker for cell proliferation in cervical cancer. Biomedical reports 1:228-230.

59. Hu J, and Gao DZ. 2012. Distinction immune genes of hepatitis-induced hepatocellular carcinoma. Bioinformatics 28:3191-3194.

60. Coutinho-Camillo CM, Salaorni S, Sarkis ÁS, and Nagai MA. 2006. Differentially expressed genes in the prostate cancer cell line LNCaP after exposure to androgen and anti-androgen. Cancer genetics and cytogenetics 166:130-138.

61. Lyng H, Brøvig RS, Svendsrud DH, Holm R, Kaalhus O, Knutstad K, Oksefjell H, Sundfør K, Kristensen GB, and Stokke T. 2006. Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer. BMC genomics 7:268.

62. Wright EM, and Turk E. 2004. The sodium/glucose cotransport family SLC5. Pflügers Archiv 447:510-518.

63. Database GCHG. SLC5A3 Gene(Protein Coding). GeneCards. Available at http://www.genecards.org/cgi-bin/carddisp.pl?gene=SLC5A3 (accessed 5th July 2017)

64. Chung YM, Kim BG, Park CS, Huh SJ, Kim J, Park JK, Cho SM, Kim BS, Kim JS, and Yoo YD. 2005. Increased expression of ICAM-3 is associated with radiation resistance in cervical cancer. International journal of cancer 117:194-201.
65. Narod S, Seth A, and Nam R. 2008. Fusion in the ETS gene family and prostate cancer. British journal of cancer 99:847-851.

66. Le S, Davis C, Konopka JB, and Sternglanz R. 1997. Two new S-phase-specific genes from Saccharomyces cerevisiae. Yeast 13:1029-1042.

67. Corpet A, De Koning L, Toedling J, Savignoni A, Berger F, Lemaitre C, O'sullivan RJ, Karlseder J, Barillot E, and Asselain B. 2011. Asf1b, the necessary Asf1 isoform for proliferation, is predictive of outcome in breast cancer. The EMBO journal 30:480-493.

68. Rosty C, Sheffer M, Tsafrir D, Stransky N, Tsafrir I, Peter M, de Cremoux P, de La Rochefordière A, Salmon R, and Dorval T. 2005. Identification of a proliferation gene cluster associated with HPV E6/E7 expression level and viral DNA load in invasive cervical carcinoma. Oncogene 24:7094-7104.

**Supplementary Information**

Appendix 1– Result of LIMMA analysis

Appendix 2– Gene expression dataset of combined dataset

Appendix 3– GSEA analysis result
Figure 1 (on next page)

Framework for the proposed integrative approach of meta-analysis and machine learning in gene expression profiling of cervical cancer.

*DE- differential expression value
Cervical cancer microarray gene expression dataset

Data Normalization

Linear Models for Microarray Analysis (LIMMA)
(select top 10000 genes from each dataset)

Intersection process

*RankProd* analysis
(select up regulated and down regulated genes)

Common genes

Potential cervical cancer genetic markers
Figure 2

Hierarchical clustering of the selected genes. (a) Hierarchical clustering of the 32 upregulated genes. (b) Hierarchical clustering of the 33 downregulated genes.

In the figures, each gene is represented by the rows and the samples are represented by the columns. The dendogram at the side indicates the relation between the pattern of the gene expression while the top-dendogram indicates the relation between the samples used. The level of expression of the genes, relative to the mean of the gene across all samples, is indicated by the color key, with the green representing the higher expression of the genes. The color bar on the left side represents the cluster of the genes after they have been cut at 1.5 of the height of the tree so that the clustering of the genes is seen more clearly. The color of the bar is indicated by the color key similarly.
Figure 3

Pathway network of the upregulated genes.
Figure 4

Pathway network of the downregulated genes.
Table 1 (on next page)

Details of the four microarray datasets used in this research.
1 Details of the four microarray datasets used in this research.

| Study                        | Platform          | Samples | Genes | Variables                                           |
|------------------------------|-------------------|---------|-------|-----------------------------------------------------|
| Noordhuis MG et al (GSE26511) [20] | HG-U133_Plus_2    | 39      | 54675 | Lymph node status: negative, positive{19}           |
| Bachtiary B et al (GSE5787) [21] | HG-U133_Plus_2    | 33      | 54675 | Biopsy: 1{5}, 2{6}, 3{7}, 4{5}, 5{5}, 6{3}, 7{2}  |
| Scotto L et al (GSE9750) [22] | HG-U133A          | 61      | 22283 | tissue: normal{24}, cell line{5}, cancer{28}       |
| den Boon JA et al (GSE63514) [23] | HG-U133_Plus_2    | 130     | 54675 | tissue: normal{24}, CIN1 lesions{14}, CIN2 lesions{22}, CIN3 lesions{40}, cancer{28} |
Table 2 (on next page)

Results of RankProd analysis- 32 Upregulated genes.
## Results of RankProd analysis - 32 Upregulated genes.

| ID_REF   | Identifier | gene.index | RP/Rsum | FC: (class1/class2) | pfp       | P.value     |
|----------|------------|------------|---------|---------------------|-----------|-------------|
| 206731   | CNKSR2     | 102        | 2.697   | 1.13                | 7.77E-155 | 6.64E-157   |
| 209243 s | PEG3       | 107        | 2.846   | 1.13                | 1.38E-149 | 2.36E-151   |
| 210549 s | CCL23      | 110        | 4.156   | 1.19                | 3.15E-114 | 8.08E-116   |
| 216108   | LOC105373738 | 22       | 4.363   | 1.22                | 3.41E-110 | 1.17E-111   |
| 220298 s | SPATA6     | 37         | 6.35    | 1.29                | 6.39E-81  | 2.73E-82    |
| 212797   | SORT1      | 20         | 6.371   | 1.33                | 9.09E-81  | 4.66E-82    |
| 207996 s | LDLRAD4    | 71         | 10.32   | 1.66                | 1.72E-49  | 1.03E-50    |
| 208762   | SUMO1      | 59         | 11.02   | 1.66                | 8.37E-46  | 5.72E-47    |
| 207257   | EPO        | 84         | 12.25   | 1.48                | 4.26E-40  | 3.28E-41    |
| 202035 s | SFRP1      | 73         | 12.53   | 1.50                | 5.99E-39  | 5.12E-40    |
| 204672 s | ANKRD6     | 86         | 13.23   | 1.57                | 3.28E-36  | 3.09E-37    |
| 220994 s | STXBP6     | 62         | 13.62   | 1.57                | 9.11E-35  | 9.34E-36    |
| 213652   | PCSK5      | 88         | 14.12   | 1.67                | 4.76E-33  | 5.29E-34    |
| 213994 s | SPON1      | 50         | 14.35   | 1.61                | 2.72E-32  | 3.25E-33    |
| 221606 s | HMGN5      | 35         | 15.39   | 1.68                | 5.13E-29  | 6.57E-30    |
| 203440   | CDH2       | 38         | 15.74   | 1.65                | 5.24E-28  | 7.17E-29    |
| 207046   | HIST2H4B   | 33         | 16.79   | 1.66                | 3.64E-25  | 5.29E-26    |
| 211865 s | FZR1       | 99         | 19.24   | 1.69                | 1.47E-19  | 2.26E-20    |
| 215401   | AU147698   | 17         | 19.79   | 1.74                | 1.72E-18  | 2.79E-19    |
| 213068   | DPT        | 4          | 20.42   | 1.75                | 2.48E-17  | 4.24E-18    |
| 214117 s | BTD        | 34         | 22.69   | 1.80                | 1.28E-13  | 2.30E-14    |
| 34187    | AK026407   | 100        | 23.65   | 1.83                | 2.78E-12  | 5.24E-13    |
| 213611   | AQP5       | 82         | 25.09   | 1.86                | 1.82E-10  | 3.58E-11    |
| 209465 x | PTN        | 53         | 26.18   | 2.00                | 3.00E-09  | 6.16E-10    |
| 208606 s | WNT4       | 13         | 27.42   | 1.92                | 5.31E-08  | 1.14E-08    |
| 211814 s | CCNE2      | 52         | 33.28   | 2.00                | 0.001144  | 0.0002542   |
| 208790 s | PTRF       | 81         | 33.49   | 2.00                | 0.00144   | 0.0003322   |
| 200908 s | RPLP2      | 40         | 33.97   | 2.04                | 0.002479  | 0.0005933   |
| 219795 s | SLC6A14    | 57         | 34.09   | 2.22                | 0.002761  | 0.0006842   |
| 205730 s | ABLIM3     | 64         | 35.48   | 2.05                | 0.01187   | 0.003044    |
| 202648 s | TCF3       | 32         | 37.13   | 2.10                | 0.05047   | 0.01337     |
| 216267 s | TMEM115    | 95         | 37.72   | 2.07                | 0.07745   | 0.02118     |
Table 3 (on next page)

Results of RankProd analysis- 33 Downregulated genes.

*Notes: Description of Table 2 and Table 3: gene.index- index of genes; RP/Rsum- computed Rank Product statistics; FC: (class1/class2)- average expression levels’ computed fold change under two conditions (upregulated class and downregulated class); pfp- estimated false positive predictions (pfp) value of the genes; and P.value- estimated p-value of each gene.
## Results of RankProd analysis - 33 Downregulated genes.

| ID_REF     | Identifier | gene.index | RP/Rsum | FC:(class1/class2) | pfp     | P.value   |
|------------|------------|------------|---------|--------------------|---------|-----------|
| 221475_s_at| RPL15      | 104        | 3.889   | 0.67               | 7.27E-120| 1.24E-121 |
| 203535_at  | S100A9     | 103        | 4.437   | 0.71               | 1.20E-108| 3.09E-110 |
| 209719_x_at| SERPINB3   | 3          | 6.49    | 0.46               | 2.94E-79 | 1.00E-80  |
| 200074_s_at| GUK1       | 55         | 7.095   | 0.45               | 4.17E-73 | 1.78E-74  |
| 209351_at  | KRT14      | 26         | 7.505   | 0.393              | 2.33E-69 | 1.20E-70  |
| 201097_s_at| SERPINB3   | 1          | 13.56   | 0.13               | 5.96E-35 | 5.60E-36  |
| 200761_s_at| ARL6IP5    | 65         | 13.75   | 0.13               | 2.67E-34 | 2.74E-35  |
| 201619_at  | PRDX3      | 69         | 14.18   | 0.10               | 7.82E-33 | 8.69E-34  |
| 206276_at  | LY6D       | 47         | 14.37   | 0.15               | 3.27E-32 | 3.91E-33  |
| 201098_at  | COPB2      | 42         | 15.65   | 0.07               | 3.17E-28 | 4.07E-29  |
| 201653_at  | CNIH1      | 87         | 16.31   | 0.067              | 2.05E-26 | 2.80E-27  |
| 211906_s_at| SERPINB4   | 8          | 16.42   | 0.003              | 3.80E-26 | 5.52E-27  |
| 202753_at  | PSMD6      | 56         | 17.6    | 0.01               | 3.54E-23 | 5.45E-24  |
| 202209_at  | LSM3       | 15         | 17.92   | 0.004              | 1.97E-22 | 3.20E-23  |
| 211023_at  | PDHB       | 9          | 17.95   | 0.018              | 2.14E-22 | 3.66E-23  |
| 221896_s_at| HIGD1A     | 5          | 19.47   | 954.6              | 3.69E-19 | 6.62E-20  |
| 213164_at  | SLC5A3     | 63         | 22.18   | 787.9              | 2.11E-14 | 3.97E-15  |
| 201863_at  | FAM32A     | 106        | 22.74   | 767.2              | 1.42E-13 | 2.78E-14  |
| 209694_at  | PTS        | 66         | 24.92   | 688.4              | 1.09E-10 | 2.23E-11  |
| 218845_at  | DUSP22     | 10         | 26.22   | 675.4              | 3.21E-09 | 6.86E-10  |
| 203315_at  | NCK2       | 80         | 27.84   | 584.5              | 1.28E-07 | 2.84E-08  |
| 213357_at  | GTF2H5     | 114        | 28.09   | 612.8              | 2.09E-07 | 4.83E-08  |
| 217850_at  | SNORD19B   | 29         | 28.27   | 592.6              | 2.95E-07 | 7.06E-08  |
| 218283_at  | SS18L2     | 12         | 28.72   | 578.6              | 6.99E-07 | 1.73E-07  |
| 203328_at  | GBE1       | 14         | 31.25   | 526.7              | 5.77E-05 | 1.48E-05  |
| 212488_at  | COL5A1     | 68         | 32.86   | 426.3              | 0.0005592 | 0.000148  |
| 219389_at  | SUSD4      | 48         | 34.36   | 439.2              | 0.003389 | 0.000927  |
| 218115_at  | ASF1B      | 36         | 35.81   | 405                | 0.00481 | 0.004177  |
| 201692_at  | SIGMAR1    | 117        | 36.53   | 393                | 0.00775 | 0.008063  |
Table 4 (on next page)

Ranking of each gene (upregulated) using proposed FS methods.

*The 9 upregulated genes selected by most (≥ 6) of the methods are CNKSR2, PEG3, SPATA6, SPON1, BTD, AK026407, RPLP2, TCF3 and TMEM115.*
## Ranking of each gene (upregulated) using proposed FS methods.

| ID_REF | Identifier | HC | PCC | Relief-F | SFS | SVM-RFE | CFS | RF | IG | Score |
|--------|------------|----|-----|----------|-----|---------|-----|----|----|-------|
| 206731 at | CNKSR2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8 |
| 209243 s at | PEG3 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 7 |
| 210549 s at | CCL23 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 5 |
| 216108 at | LOC105373738 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 5 |
| 220298 s at | SPATA6 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 7 |
| 212797 at | SORT1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 4 |
| 207996 s at | LDLRAD4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 208762 at | SUMO1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 3 |
| 207257 at | EPO | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 5 |
| 202035 s at | SFRP1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 4 |
| 204672 s at | ANKRD6 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 5 |
| 220994 s at | STXB6 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 5 |
| 213652 at | PCSK5 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 3 |
| 213994 s at | SPON1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 6 |
| 221606 s at | HMGN5 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 203440 at | CDH2 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 5 |
| 207046 at | HIST2H4B | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 211865 s at | FZR1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 5 |
| 215401 at | AU147698 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 5 |
| 213068 at | DPT | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 5 |
| 214117 s at | BTD | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 7 |
| 34187 at | AK026407 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 6 |
| 213611 at | AQP5 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 5 |
| 209465 s at | PTN | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 3 |
| 208606 s at | WNT4 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 5 |
| 211814 s at | CCNE2 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| 208790 s at | PTRF | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 200908 s at | RPLP2 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 7 |
| 219795 at | SLC6A14 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 |
| 205730 s at | ABLIM3 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 5 |
| 202548 at | TCF3 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 7 |
| 216267 s at | TMEM115 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 6 |
**Table 5** (on next page)

Ranking of each gene (downregulated) using proposed FS methods.

*The 12 downregulated genes selected by most (≥ 6) of the methods are RPL15, HIGD1A, ARL6IP5, PRDX3, COPB2, CNIH1, PSMD6, LSM3, PDHB, SLC5A3, PTS and ASF1B. As gene HIGD1A is selected twice with different ID REF, it only considered once here.*
Ranking of each gene (downregulated) using proposed FS methods.

| ID_REF       | Identifier | HC | PCC | Relief-F | SFS | SVM-RFE | CFS | RF | IG | Score |
|--------------|------------|----|-----|----------|-----|---------|-----|----|----|-------|
| 221475_s_at  | RPL15      | 1  | 1   | 1        | 0   | 1       | 1   | 1  | 1  | 7     |
| 203535       | SI00A9     | 0  | 0   | 1        | 1   | 1       | 1   | 0  | 4  | 2     |
| 209719_x_at  | SERPINB3   | 0  | 0   | 0        | 1   | 1       | 0   | 0  | 2  | 0     |
| 200074_s_at  | GUK1       | 1  | 0   | 0        | 0   | 0       | 1   | 0  | 3  | 2     |
| 209351       | KRT14      | 0  | 0   | 0        | 0   | 1       | 0   | 0  | 2  | 7     |
| 201097_s_at  | ARF4       | 1  | 1   | 0        | 1   | 1       | 1   | 0  | 4  | 0     |
| 217845_x_at  | HIGD1A     | 1  | 1   | 1        | 1   | 1       | 1   | 0  | 4  | 7     |
| 210413_x_at  | SERPINB4   | 0  | 1   | 0        | 0   | 1       | 1   | 1  | 0  | 4     |
| 210835_s_at  | CTBP2      | 1  | 0   | 0        | 0   | 1       | 1   | 1  | 4  |       |
| 209720_s_at  | SERPINB3   | 0  | 1   | 0        | 1   | 1       | 0   | 1  | 3  | 3     |
| 200761_s_at  | ARL6IP5    | 1  | 1   | 1        | 1   | 1       | 0   | 1  | 0  | 6     |
| 201619       | PRDX3      | 1  | 1   | 1        | 1   | 1       | 0   | 0  | 7  |       |
| 206276       | LY6D       | 0  | 0   | 0        | 1   | 1       | 1   | 0  | 3  | 1     |
| 201098       | COPB2      | 1  | 1   | 1        | 1   | 1       | 1   | 1  | 7  | 3     |
| 201653       | CNIH1      | 1  | 1   | 1        | 0   | 1       | 1   | 1  | 7  |       |
| 211906_s_at  | SERPINB4   | 0  | 1   | 0        | 1   | 0       | 1   | 0  | 3  | 3     |
| 202753       | PSMD6      | 1  | 1   | 0        | 1   | 0       | 1   | 1  | 6  |       |
| 202209       | LSM3       | 1  | 1   | 1        | 1   | 1       | 0   | 0  | 6  |       |
| 211023       | PDHB       | 1  | 1   | 1        | 1   | 1       | 1   | 1  | 8  |       |
| 221896_s_at  | HIGD1A     | 1  | 1   | 1        | 1   | 1       | 0   | 1  | 0  | 6     |
| 213164       | SLC5A3     | 1  | 0   | 0        | 1   | 1       | 1   | 1  | 6  |       |
| 201863       | FAM32A     | 1  | 0   | 1        | 1   | 0       | 0   | 0  | 4  |       |
| 209694       | PTS        | 1  | 1   | 1        | 1   | 1       | 1   | 0  | 0  | 6     |
| 218845       | DUSP22     | 1  | 0   | 0        | 1   | 0       | 1   | 1  | 5  |       |
| 203315       | NCK2       | 1  | 0   | 0        | 1   | 1       | 0   | 1  | 5  |       |
| 213357       | GTF2H5     | 1  | 1   | 0        | 0   | 0       | 0   | 0  | 4  | 3     |
| 217850       | SNORD19B   | 1  | 1   | 0        | 0   | 1       | 0   | 1  | 5  | 2     |
| 218283       | SS18L2     | 1  | 0   | 1        | 0   | 0       | 0   | 0  | 3  | 4     |
| 203282       | GBE1       | 1  | 1   | 0        | 0   | 0       | 0   | 0  | 3  |       |
| 212488       | COL5A1     | 1  | 0   | 1        | 0   | 0       | 0   | 0  | 3  |       |
| 219389       | SUSD4      | 1  | 0   | 0        | 0   | 0       | 1   | 0  | 4  | 7     |
| 218115       | ASFB1      | 1  | 0   | 1        | 1   | 1       | 1   | 1  | 7  |       |
| 201692       | SIGMAR1    | 1  | 0   | 1        | 0   | 0       | 1   | 1  | 5  |       |
Table 6 (on next page)

Results of GSEA analysis using the 9 upregulated-genes.

*There was no matching of gene AK026407, TCF3 and TMEM115 with the MSigDB.
1 Results of GSEA analysis using the 9 upregulated-genes.

| Identifier | Highest Normalized Enrichment Score (NES) | FDR q-value | Status   |
|------------|------------------------------------------|-------------|----------|
| CNKSR2     | 1.00                                     | 0.859       | Rejected |
| PEG3       | 1.33                                     | 0.000       | Pass     |
| SPATA6     | 1.00                                     | 0.823       | Rejected |
| SPON1      | 1.60                                     | 0.053       | Pass     |
| BTD        | 1.50                                     | 0.026       | Pass     |
| AK026407   | -                                        | -           | -        |
| RPLP2      | 1.40                                     | 0.000       | Pass     |
| TCF3       | -                                        | -           | -        |
| TMEM115    | -                                        | -           | -        |
**Table 7** (on next page)

Results of GSEA analysis using the 12 downregulated-genes.

*There was no matching of gene ARL6IP5, CNIH1 and PTS with the MSigDB.*
## Results of GSEA analysis using the 12 downregulated-genes.

| Identifier | Highest Normalized Enrichment Score (ES) | FDR q-value | Status |
|------------|------------------------------------------|-------------|--------|
| RPL15      | 1.21                                     | 0.779       | Rejected |
| HIGD1A     | 1.25                                     | 1.00        | Rejected |
| ARL6IP5    | -                                        | -           | -      |
| PRDX3      | 1.41                                     | 0.020       | Pass   |
| COPB2      | 1.41                                     | 0.020       | Pass   |
| CNIH1      | -                                        | -           | -      |
| PSMD6      | 1.44                                     | 0.560       | Rejected |
| LSM3       | 1.41                                     | 0.020       | Pass   |
| PDHB       | 1.33                                     | 0.839       | Rejected |
| SLC5A3     | 1.50                                     | 0.000       | Pass   |
| PTS        | -                                        | -           | -      |
| ASF1B      | 1.43                                     | 0.000       | Pass   |
Table 8 (on next page)

Function, related diseases and description of the most significant genes (Upregulated).
| GENES   | FULL NAME                                      | FUNCTION                                           | RELATED DISEASES AND DESCRIPTION                                      |
|---------|-----------------------------------------------|----------------------------------------------------|-----------------------------------------------------------------------|
| PEG3    | Paternally Expressed Gene 3                   | protein coding gene.                                | • Hypermethylation of PEG3 is linked to epigenetics mechanism and HPV infection in cervical intraepithelial neoplasia (CIN) and Invasive Cervical Cancer. [40]  |
|         |                                               |                                                    | • Cell proliferation and p53-mediated apoptosis. [41]                                                                 |
|         |                                               |                                                    | • Involvement in the glioma and ovarian cancers. [42]                                                              |
| SPON1   | F-spondin 1                                   | extracellular matrix organization regulation, interaction between cells and axon guidance. [43] | • Showed extreme expression activities in the identification of colorectal biomarkers. [44]  |
|         |                                               |                                                    | • Down regulation activity in the cervical cell carcinoma. [45]                                                      |
| BTD     | bitonidase                                    | gene expression, proliferation and differentiation of cells, gene signaling. [46] | • gene expression signature that marks pelvic lymph node metastasis (PLNM) in cervical carcinoma. [47]  |
|         |                                               |                                                    | • Involvement in thyroid and lung cancers [48, 49]                                                                 |
|         |                                               |                                                    | • Potential serological marker in breast cancer plasma. [50]                                                          |
| RPLP2   | Ribosomal Protein Lateral Stalk Subunit P2    | encodes a 60s subunit ribosomal protein.            | • prognostic marker for gastric cancer. [51]                                                                         |
|         |                                               |                                                    | • involved in the carcinogenesis and progression of various cancers. [52]                                              |
|         |                                               |                                                    | • associated with other ribosomal protein genes with colorectal carcinomas (CRC). [53]                              |
Table 9 (on next page)

Function, related diseases and description of the most significant genes (Downregulated).
1. **Function, related diseases and description of the most significant genes (Downregulated).**

| GENES | FULL NAME | FUNCTION | RELATED DISEASES AND DESCRIPTION |
|-------|-----------|----------|----------------------------------|
| **PRDX3** | Peroxiredoxin3 | • encodes antioxidant function protein, provides protection to mitochondria from oxidative stress. [55]  
• regulates the cellular Reactive Oxidative Species (ROS) level of the cell. [56] | • showed correlation with the severity of the cervical carcinoma [57].  
• Showed immunopositivity to be significantly higher in cervical cancer cells [58] |
| **COPB2** | Coatomer Protein Complex Subunit Beta 2 | • protein coding gene.  
• involves in the transduction of signal for G protein. [59] | • One of the differentially expressed genes in LNCaP prostate cancer cell lines. [60] |
| **LSM3** | Hypothetical protein LOC285378 | • involves in the process of mRNA splicing | • Involved in the rapid proliferation, invasiveness, oxidative phosphorylation and tumor size of cervical cancers. [61] |
| **SLC5A3** | Solute Carrier Family 5 Member 3 | • involves in the osmoregulation of cells. [62]  
• prevents the impairment of cell cellular function [63] | • differentially expressed between Parental SiHa Cells and SiHa/R Cells of cervical cancer. [64]  
• involved in the formation of fusion transcripts gene in some prostate cancer cases. [65] |
| **ASF1B** | Anti-Silencing Function 1B Histone Chaperone | • codes for the substrate protein of the cycle regulated-kinase cell.  
• derepress the overexpression of transcriptional silencing. [66] | • associated with the aggressiveness of breast tumor. [67]  
• led to the poor disease outcome in the study of the cervical cancer proliferation cluster that corresponding to the 163 transcripts [68]. |