**Gene Expression Profile of Ovarian Cancer Dataset that helps to Predict Potential Biomarkers**

Divya, V., Anitha P. Muttagi, Seema J Patel, Gurumurthy, H. and Prashantha Nagaraja

GM Institute of Technology, Department of Biotechnology, Davanagere, Bangalore-560092

**ABSTRACT**

The viral pathogens such as human papillomavirus (HPV), cytomegalovirus (CMV) and Chlamydia trachomatis are significant risk factors for developing women mucinous epithelial ovarian cancer. The clinical and morphological distinct of ovarian cancer subtypes of frequent concurrence of endometriosis results in poor prognosis. The lack of significant markers associated with diagnosis in early stage of infection. The aim of the study is conducted oligonucleotide microarray to compare image analysis and normalization algorithms to analyze host pathogen interacting genes and proteins that reside on network of disease susceptibility. The expression is calculated on the ratio between virus-infected tissues and normal tissues are brought great expectations for finding biomarkers that would improve patient’s treatment in the early stage of infection. The intensity of gene chip is processed and normalization is carried out using R statistical software. We have identified significant clusters of highly enriched gene markers that extent in epithelial malignancies and predicted the large expression data of candidate molecular biomarkers.

**Key words:** ovarian cancer, viral cancers, human papilloma virus, cytomegalovirus, Microarray, image analysis, statistical analysis, hierarchical clustering, k-means clustering

**INTRODUCTION**

Ovarian cancer is a most leading cause of death from gynecological malignancy in all over the world (Greenlee et al., 2001). Approximately, 70% of women are diagnosed with ovarian adenocarcinoma are derived from ovarian surface epithelia (OSE) (Sevely et al., 2004; Auerberg et al., 2001). Based on histological subtypes of epithelial ovarian carcinomas shows 40% of population is affected with viral infection such as human papillomavirus (HPV), Chlamydia trachomatis and cytomegalovirus (CMV) (Atalay et al., 2007). The closer look at HPV virus infected with cervical cancer is link with ovarian cancer is still unclear (HPV, 2011). The women with sexually active has infected with HPV family members HPV-16 and HPV-18. The high risk of HPV infected subtypes produce two types of oncogenes designed E6 and E7 proteins induce transformation by interference with endogenous cell cycle regulatory proteins, including P53, retinoblastoma (Rb) and breast cancer type 1 susceptibility protein (BRAC1) (Fishman et al., 2002). The HPV virus is clearly link with different cancers such as head and neck cancers (Giordano et al., 2008), still there is lot of investigations is going on to identify the genes, functions and pathways that helps to identify potential drug targets. The virus such as Chlamydia trachomatis is affected with women having sexually transmitted disease (Claman et al., 1997).

There are several serotypes that may cause urogenital infection which evolve in chronic infection results in spreading to female pelvic blockage to cause infertility (Den Hartog, 2005; Dieterle and Wollenhaupt, 1996). The pathogenesis of the c.trachomatis is unknown, but the bacterium may affect non immune cells that produce proinflammatory response against host cell (Idahl et al., 2007; Tiitinen et al., 2006). The genomic function and structure activity remains unknown, the bacterium express high levels of Chlamydial heat shock protein 60 (cHSP60) affects immune system and cause tubal factor infertility (Madeleine et al., 2007). There are several other pathogens such as mycobacterium tuberculosis is also associated with squamous cell carcinoma of the cervix (Koskela et al., 2000).

The role of persistent infection, leading to chronic inflammation, in the pathogenesis of ovarian cancer has received very little consideration, although a history of pelvic inflammatory disease (PID) is in a case-control study correlated to higher risk for ovarian cancer (Risch et al., 1995). Recent histopathological studies of different genomes involved in disease causing are a major task in disease identification. There are different histotypes helps to identify the differently expressed genes that significantly associated with pathogenesis in both cancerous and non-cancerous tissues that is suitable prediction of drug identification or vaccines preparation to cure the disease.
MATERIALS AND METHODS

Raw Data selection and pre-processing of microarray data

In order to determine the pathogenic genes involved in ovarian cancer and the role of gene expression in disease progression can be analyzed using microarray data. The Comparative gene expression profiling analyses were carried out to determine disease mechanism and the role of signaling pathways, (i) Gene expression measurements (ii) definitions of signaling pathways and (iii) protein drug targets prediction. We have evaluated all published case control studies and diseased datasets were selected using various repositories such as GEO (Gene expression Omnibus), Array express (EBI database), and PUMAdb (Princeton University Microarray database). In order to determine the Chlamydia trachomatis in ovarian cancer dataset such as GSE41075 (Vicetti Miguel et al., 2013) and Human Papilloma virus infected dataset GSE49288 (Kim and Shin, 2013) is used for studies.

The GSE41075 dataset contains 22 samples of which 10 controls trans-cervical, endometrial biopsy specimens of upper and lower genital tract infection and 12 women of C. Trachomatis endometrial infection, cells were processed for microarray analysis using Affymetrix Human genome U133A Gene Chip. The GSE49288 dataset has 39 cervical cancer samples were grouped into 4 sets based on physical examination and differential expression of gene profiles using Agilent two-color experiment. The Affymetrix and Agilent datasets is pre-processed using RMA and MAS5 algorithms. However, the signal intensity of MM probe can often be larger than the PM probe implying that MM probe is detecting a true signal as well as background signal. Probe set results were further evaluated using R and BioConductor software Probes were considered differentially expressed if they had a fold change value of ≥ 3 and a p-value < .005 (Student's t-test). The sequence clusters were created from the UniGene database and then refined by analysis and comparison with a number of other publicly available databases.

Identification of Differential gene expression data analysis

The preprocessed dataset is used for differential gene expression studies using limma package. The RMA function assigned the factorial design expression that transforming log2 values. To assign column names of eset creates contrast matrix to perform all pairwise comparisons to compute estimated coefficients and standard errors of a given datasets. Computes moderated t-statistics and log-odds of differential expression by empirical Bayes shrinkage of the standard errors towards a common value. Generates list of top 10 (‘number=10’) differentially expressed genes sorted by B-values (‘sort.by=B’) for each of the three comparison groups (‘coef=1’) in this sample set. The summary table has logFC is the log2 for each of the three comparison groups (‘coef=1’) in this sample set. The differential gene expression is predicted based on log2 fold changes, standard errors, t-statistics and p-values. The analysis of significance is based on probe intensity that is interpreted by ordinary t-statistic except that the standard errors have been moderated across genes, i.e., shrunk towards a common value, using a simple Bayesian model. Using functional annotation, we have predicted 1643 differentially expressed genes, out of these 276 up regulated and 1367 down regulated genes expressed in Chlamydia Trachomatis infection in ovarian cancer. There are 158 genes is significantly associated with protein expression of which 26 protein functional genes is used for potential drug targets. The network prediction of differently expressed genes such as RPL24, CYP2A6, CYP2E of Cytochrome P45 subfamily, CST3, DDX5, RPS27A, NONO, JUN, TP53, EGFR1, NCOR1, HNRNP1 and EGFR genes is predicted in subnetworks of functional annotated and enriched data by DAVID to reveal over-represented biological functions. The gene expression signatures is calculated the distance matrix for the disease pairs based on the overlap between sets of differentially expressed genes used for potential drug targets.

RESULTS

The pre-processed data of Chlamydia Trachomatis is used for differential gene expression analysis. There are 24789 genes in a dataset is and used for pair wise comparisons with different series of samples. The differential gene expression is predicted based on log2 fold changes, standard errors, t-statistics and p-values. The analysis of significance is based on probe intensity that is interpreted by ordinary t-statistic except that the standard errors have been moderated across genes, i.e., shrunk towards a common value, using a simple Bayesian model. Using functional annotation, we have predicted 1643 differentially expressed genes, out of these 276 up regulated and 1367 down regulated genes expressed in Chlamydia Trachomatis infection in ovarian cancer. There are 158 genes is significantly associated with protein expression of which 26 protein functional genes is used for potential drug targets. The network prediction of differently expressed genes such as RPL24, CYP2A6, CYP2E of Cytochrome P45 subfamily, CST3, DDX5, RPS27A, NONO, JUN, TP53, EGFR1, NCOR1, HNRNP1 and EGFR genes is predicted in subnetworks of close association and is used for potential drug targets, the overall results is predicted in Table: 1, Figure: 1a-h.

The HPV infected dataset is preprocessed using Agilent platform of RMA and MAS5 algorithms that used for differential gene expression analysis. The overall significant differences of gene signatures were listed in table 2a-d. We
Table 1. Significant gene signatures is predicted from Chlamydia Trachomatis is differentially expressed in ovarian and cervical cancer

<table>
<thead>
<tr>
<th>Description</th>
<th>P-value</th>
<th>FDR q-value</th>
<th>Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear gene expression</td>
<td>1.68E-6</td>
<td>2.18E-60</td>
<td>3.67</td>
</tr>
<tr>
<td>SRP-dependent cotranslational protein targeting to membrane</td>
<td>3.26E-5</td>
<td>2.11E-50</td>
<td>11.84</td>
</tr>
<tr>
<td>cotranslational protein targeting to membrane</td>
<td>2.01E-5</td>
<td>8.69E-50</td>
<td>11.62</td>
</tr>
<tr>
<td>protein targeting to ER</td>
<td>4.46E-5</td>
<td>1.45E-49</td>
<td>15.9</td>
</tr>
<tr>
<td>protein localization to endoplasmic reticulum</td>
<td>4.51E-5</td>
<td>1.71E-49</td>
<td>10.65</td>
</tr>
<tr>
<td>establishment of protein localization to endoplasmic reticulum</td>
<td>3.77E-5</td>
<td>8.16E-49</td>
<td>15.46</td>
</tr>
<tr>
<td>protein targeting to membrane</td>
<td>1.15E-4</td>
<td>2.13E-46</td>
<td>12.49</td>
</tr>
<tr>
<td>viral process</td>
<td>3.89E-4</td>
<td>6.30E-46</td>
<td>3.24</td>
</tr>
<tr>
<td>symbiosis, encompassing mutualism through parasitism</td>
<td>3.89E-4</td>
<td>5.60E-46</td>
<td>3.24</td>
</tr>
<tr>
<td>multi-organism cellular process</td>
<td>1.63E-4</td>
<td>2.12E-45</td>
<td>3.21</td>
</tr>
<tr>
<td>interspecies interaction between organisms</td>
<td>2.34E-4</td>
<td>2.77E-44</td>
<td>3.01</td>
</tr>
<tr>
<td>establishment of protein localization to membrane</td>
<td>2.79E-4</td>
<td>3.02E-42</td>
<td>6.44</td>
</tr>
<tr>
<td>nuclear-transcribed mRNA catabolic process, nonsense-mediated decay</td>
<td>6.17E-4</td>
<td>6.16E-42</td>
<td>15.11</td>
</tr>
<tr>
<td>Translation</td>
<td>6.22E-4</td>
<td>5.76E-42</td>
<td>7.26</td>
</tr>
<tr>
<td>viral transcription</td>
<td>1.97E-4</td>
<td>1.71E-40</td>
<td>15.31</td>
</tr>
<tr>
<td>mRNA metabolic process</td>
<td>4.26E-4</td>
<td>3.45E-40</td>
<td>4.32</td>
</tr>
<tr>
<td>protein targeting</td>
<td>4.14E-4</td>
<td>3.16E-38</td>
<td>7.79</td>
</tr>
</tbody>
</table>

Fig. 1a-h. Differentially expressed gene signatures in protein functions of ovarian cancer
Site of expression for Ovarian cancer

Transcription factor for Ovarian cancer

Clinical phenotypes for Ovarian cancer
have classified 4 different groups which are top ranked with functional annotation and enrichment studies. We have identified 19848 upregulated, 1860 down regulated genes that significantly expressed in cervical tissues.

Using significant association studies shows only 100 genes that eventually predicts functional characters of proteins. We have predicted AU146383 probe is highly expressed HMEC-1 endothelial cell line response of transcriptional factor-induced pluripotent stem cells of ovarian tissues. The other probe such as AF130077, IGFBP1, AFP, AGT, ABC2, ITH3, ApoB and ApoH genes is mainly involved in HPV transcriptional regulation and immune suppression on host cells. These selected proteins are mainly used for potential drug targets against cervical and ovarian cancer.

**Conclusion**

On ongoing challenges of drug targets identification of different cancer types based on type of pathogenesis, disease progression, risk factors and signs and symptoms is evidence to slow down the development of novel anticancer agents. We have used computational methods to identify drug targets based on gene expression of ovarian and cervical cancer associated pathogens such as Chlamydia trachomatis and HPV. Using different statistical analysis to identify differential expressions of genes involved in ovarian and cervical cancer furthermore used genome-wide scale of functional annotation and functional enrichment studies to identify potential drug targets based on protein expression. We have identified the top level expressed genes such as RPL24, CYP2A6, and CYP2E
of Cytochrome P45 subfamily, CST3, DDX5, RPS27A, NONO, JUN, TP53, EGR1, NCO1, HNRNPU and EGFR is mainly targeted to Chlamydia trachomatis infection in cervical and ovarian cancer infection. The HPV targeted genes such as AU146383, AF130077, IGFBP1, AFP, AGT, ABCC2, ITIH3, ApoB and ApoH is best drug targets against cervical and ovarian cancer tissues. We believe that the application of our integrated approach has the potential to provide a list of drug target candidates for other human diseases.

Acknowledgement

We gratefully acknowledge the assistance of Dr. C. N Prashantha for supporting project work, workflows, technical assistance and report preparation. We also grateful to Seema J Patel and Anitha P.Muttagi as internal guide to conduct seminars and guiding project. We also grateful to Department of Biotechnology, GMIT and Scientific Bio-Minds to providing continuous support to successfully complete the project.

Conflict of interest

There is no conflict of interest

REFERENCES
