An Integrative Genomics Approach to Discovery of Molecular Markers in Ovarian Cancer

Dahlia Siprut, Ms. Aditi Kuchi, Dr. Jiande Wu, Dr. Chindo Hicks

Department of Genetics, Louisiana State University Health Sciences Center, New Orleans, LA

Introduction

Ovarian cancer (OC) is one of the most common gynecologic cancers that has the highest mortality rate. OC is the eighth most commonly occurring cancer in women and the 18th most commonly occurring cancer overall. World-wide there were nearly 300,000 new cases of OC in 2018. Within the United States an estimated 22,530 women were diagnosed with new cases of OC and an estimated 13,980 women died from the disease in 2019. Therefore, there is an urgent need for the discovery of molecular markers for early detection and prognostic prediction of the disease. Advances in next generation sequencing technology have enabled generation of vast amounts of gene expression and somatic mutation data on cancer genomes including OC. Although much progress has been made on classification of molecular subtypes of OC using transcription profiling, gene expression has not been leveraged and integrated with somatic mutations information for the discovery of diagnostic and prognostic markers. The objective of this investigation was to discover prognostic markers that are predictive of clinical outcome using gene expression and somatic mutation data. Our working hypothesis was that genomic alterations in the transcriptomes and tumor genomes of women diagnosed with OC could lead to measurable changes distinguishing patients who survived the disease from those who did not survive the disease.

Materials and Methods

We addressed this hypothesis using gene expression and somatic mutation data derived from a total of 376 samples (230 died from the disease and 146 survived the disease) from the Cancer Genome Atlas (TCGA). We downloaded the information directly from the website and extracted the data into separate files using WinRAR. Once extracted, the data was compiled into two matrix, one for gene expression and one for somatic mutations. The somatic mutations were merged using a Perl Script command. A total number of each mutation (SNP, INS, DEL) was tallied and placed into columns and the "NULLs" removed. The merged data represented samples as columns and genes as rows. The information was then sorted from least expressed to most expressed and those that fell below 100 were removed from the data set. The data was partitioned into two patient groups, those who survived the disease and those who died from the disease. We performed analysis comparing gene expression levels between the two patient groups using Pomelo II to discover a signature of significantly differentially expressed genes distinguishing those that survived to those that did not. The most expressed genes, top 16,000, were run through the program using a (limma) t-test. Significantly differentially expressed genes (P<0.001) were evaluated for the presence of somatic mutations to identify a signature of significantly differentially expressed genes which were also significantly differentially mutated distinguishing the two patient groups.

Table 1

Table 1 represents the 50 significantly differentially expressed genes with P<0.001.

Gene Names:	Expression p-values:
GBP1P1	5.00E-07
CXCL11	4.00E-06
PLA2G2D	4.70E-06
TAP1	5.30E-06
CD38	9.00E-06
IGKV4-1	1.49E-05
IGHG1	1.61E-05
GBP4	2.03E-05
IGKC	2.30E-05
CXCL9	2.58E-05
SLAMF7	4.70E-05
IL18BP	6.04E-05
CXCL10	6.85E-05
MZB1	7.65E-05
GBP5	9.40E-05
CD79A	0.000104
PRRT1	0.000131
CD27	0.0001528
IGLC3	0.0001676
ETV7	0.0001714
CD2	0.0001845
SIT1	0.0001906
WARS	0.0002255
JCHAIN	0.0002301
HLA-DOB	0.0002319
TAP2	0.0002354
CPNE5	0.0002511
IL2RG	0.0002724
CXCR3	0.0003044
CD3D	0.0003062
IGHV3-30	0.0003632
TRAC	0.0003709
CD3G	0.0003717
IKZF3	0.0003938
CXCL13	0.0003964
IRF4	0.0004501
BTN3A1	0.0004584
TRBC2	0.0004693
PSMB9	0.0005044
SLAMF6	0.000542
HLA-F	0.0005791
EMP1	0.0006333
IGHV3-7	0.0006422
CCR2	0.000648
GBP1	0.0007441
HCP5	0.0007997
CD3E	0.0008264
ITK	0.000836
ADGRG5	0.0008636
CCL8	0.0008998

Table 2

Table 2 represents only the 23 significantly differentially expressed genes (P<0.001) that also contained somatic mutations.

Gene Names:	Expression p-values:	# Mutations:
TAP1	5.30E-06	11
CD38	9.00E-06	2
IGKV4-1	1.49E-05	1
CXCL9	2.58E-05	1
SLAMF7	4.70E-05	1
GBP5	9.40E-05	1
CD79A	0.000104	3
CD27	0.0001528	2
ETV7	0.0001714	1
CD2	0.0001845	26
WARS	0.0002255	1
TAP2	0.0002354	6
CXCR3	0.0003044	2
CD3D	0.0003062	1
IKZF3	0.0003938	2
CXCL13	0.0003964	1
IRF4	0.0004501	2
BTN3A1	0.0004584	1
SLAMF6	0.000542	1
EMP1	0.0006333	5
GBP1	0.0007441	1
CD3E	0.0008264	4
ITK	0.000836	7

Results

In total, 44,734 genes were differentially expressed with an expression level value over 100 and 97 genes contained mutations. Of the expressed genes, the top 16,000 were run through Pomelo II to determine those that were significantly differentially expressed. The analysis revealed a signature of 130 differentially expressed genes (P<0.005) of which 50 were significantly (P<0.001) differentially expressed genes distinguishing patients who survived from patients who died. Evaluation of these genes for the presence of somatic mutations revealed a signature of 23 significantly differentially expressed genes which were also differentially mutated distinguishing the two patient groups. Among the top somatic mutated differentially expressed genes distinguishing the two patient groups included the genes: TAP1, CD79A, CD2, TAP2, EMP1, CD3E, and ITK.

Conclusions

We discovered a signature of somatic mutated genes were differentially expressed distinguishing patients who survived OC from patients who died from the disease. Our investigation demonstrates that integrative analysis combining gene expression with somatic mutation data is a powerful approach to discovery of molecular markers predictive of clinical outcomes and clinical endpoints.. Moving forward, an analysis for differential expression of the 97 genes that contained mutation data originally may provide important data on the relationship between the two. More research into the 7 highest expressed genes containing mutations should also be considered for future projects.

This research project was supported by grant # 1659752 through the National Science Foundation (NSF), Research Experiences for Undergraduates (REU) Program , Bioinformatics and Genetics Program