Integrative Genomics Approach to Biomarker Discovery in Colorectal Cancer

Eric R. Espinal, Aditi S. Kuchi, Dr. Jiande Wu, Dr. Chindo Hicks

Bioinformatics and Genetics Program, Department of Genetics, LSUHSC, New Orleans, LA

Introduction

Despite remarkable progress in screening and patient management, colorectal cancer (CC) remains a major public health problem. CC is the third most commonly occurring cancer in men and the second most commonly occurring cancer in women. World-wide there were over 1.8 million new cases of CC in 2018. In United States it is estimated that there will be 147,950 newly diagnosed cases of CC and an estimated 53,200 individuals will die from the disease in 2020. Therefore, a critical unmet and urgent medical need is discovery of molecular markers for early detection of the disease. The recent surge of next generation sequencing technology have enabled generation of vast amounts of gene expression and somatic mutation data on CC. These advances have enabled molecular classification of subtypes and increased our understanding of the molecular taxonomy of CC. However, gene expression has not been optimally leveraged and integrated with somatic mutation information for the discovery of diagnostic markers. The objective of this investigation was to discover clinically actionable biomarkers for diagnosis and prognosis of CC using gene expression and somatic mutation data. Our working hypothesis was that genomic alterations in individuals diagnosed with CC and control samples could lead to measurable changes distinguishing patients diagnosed with CC from controls.

Methods and Materials

We addressed this hypothesis using gene expression and somatic mutation data derived from a total of 523 samples (481 CC samples and 42 control samples) from the Cancer Genome Atlas (TCGA). We extracted the data from the website using WinRAR. The data was merged to form two data matrixes: a gene expression matrix and somatic mutation matrix. To merge the mutation information, we used a command from Perl script. Once in a big file, the "NULL" spaces were deleted, and the three types of mutations (SNP, DEL, and INS) were totaled in a separate column. The data information was merged manually with each column representing a sample and each row representing a gene. At the end, we totaled the genes in each row and sorted the data from least expressed to most expressed genes. The genes with less than 100 were deleted. The data was partitioned into two data sets tumor samples and control normal. We performed analysis comparing gene expression levels between the two sample groups using Pomelo II to discover a signature of significantly differentially expressed genes distinguishing tumors from controls. The top 16,000 genes were inputted into the program and performed with a t-test (limma). Significantly differentially expressed genes (100 genes) 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 sample groups.

Table 1A and B

Gene Name	unadj.p
MT-CO1	0.0000001
MT-RNR2	0.0000001
MT-ND1	0.0000001
IGHA1	0.0000001
IGKC	0.0000001
IGHA2	0.0000001
TMBIM6	0.0000001
ATP5F1B	0.0000001
BSG	0.0000001
PFN1	0.0000001
JCHAIN	0.0000001
ITM2C	0.0000001
FXYD3	0.0000001
ATP5F1A	0.0000001
FTH1	0.0000001
S100A10	0.0000001
SERF2	0.0000001
MISP	0.0000001
CALM3	0.0000001
AES	0.0000001
DAZAP2	0.0000001
PTTG1IP	0.0000001
ATP5MC3	0.0000001
AOC1	0.0000001
CYCS	0.0000001
FCGRT	0.0000001
CTDSP2	0.0000001
COX7C	0.0000001
ARF3	0.0000001

Table 1A depicts the significantly differentially expressed genes that did not have any somatic mutations. It also shows the p-value of the genes: p<1.00x10-7.

Gene Name	unadj.p	Total Mutations
PIGR	0.0000001	6
ATP1A1	0.0000001	8
FCGBP	0.0000001	24
MYH11	0.0000001	25
MUC13	0.0000001	6
CDH17	0.0000001	10
MYH14	0.0000001	12
AHNAK	0.0000001	18
CKB	0.0000001	8
FLNB	0.0000001	14
GRN	0.0000001	5
CSDE1	0.0000001	11
PTPRF	0.0000001	11
CLCA1	0.0000001	8
CTNNA1	0.0000001	11
TLN1	0.0000001	10
IQGAP1	0.0000001	11
SLC44A4	0.0000001	5
A2M	0.0000001	12
ANPEP	0.0000001	10
CTSA	0.0000001	5
CLSTN1	0.0000001	8
HADHA	0.0000001	5
CAPN2	0.0000001	8
CALD1	0.0000001	13
RTN4	0.0000001	7
SCARB2	0.0000001	7
PLS1	0.0000001	6
ACTA2	0.0000001	4

Table 1B shows just 30 of the top genes that had somatic mutations. The total number of mutations throughout all samples is also displayed. Once again it shows the significance of the genes (p<1.00x10-7).

Results

After filtering out the non expressive genes, in total 26,397 genes were differentially expressed because they had expression levels that were over 100. We ran the top 16,000 genes through Pomelo II to determine which genes were significantly differentially expressed. The analysis revealed a signature of 100 highly significantly (p<1.00x10-7) differentially expressed genes distinguishing individuals with CC from controls. Evaluation of these genes for the presence of somatic mutations revealed a signature of 80 significantly differentially expressed genes which were also differentially mutated distinguishing the two sample groups. Among the top somatic mutated differentially expressed genes distinguishing the two samples groups included the genes ATP1A1, PIGR, FCGBP, MYH11, PTPRF, CDH17, MYH14, AHNAK, FLNB, and CSDE1. Overall, there were 16,013 genes that contained mutations. It is unclear, however, how many of these mutated genes were also differentially expressed.

Conclusions

We discovered a signature of somatic mutated genes which were differentially expressed genes distinguishing patients diagnosed with CC from controls. The majority of the differentially expressed genes also contained higher somatic mutations. Our investigation demonstrates that integrative analysis combining gene expression with somatic mutation data is a powerful approach to discovery of molecular diagnostic and prognostic markers in CC. Future research can include studying which mutated genes are driver genes for CC and which are passenger genes.

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