The Center for Research Informatics (CRI)
Biomedical Informatics Resources

Julie Johnson, MPH, RN
October 27, 2014
Bob Grossman, PhD

Sam Volchenboum, MD, PhD
Center for Research Informatics

Bioinformatics

Clinical Research Data Warehouse

Applications Development

IT Operations and Infrastructure
Center for Research Informatics

Bioinformatics

Clinical Research Data Warehouse

Applications Development

IT Operations and Infrastructure

- 8 full-time PhDs
- Industry-grade pipelines
- Custom workflows
- Grant preparation
- Manuscript writing
Project Report

Identification of Potential Prognostic Biomarkers for Ewing Sarcoma Patients

Center for Research Informatics (CRI), The University of Chicago

Project requirement: Broad Institute Genomics Center/W2000 project, Analysis and annotation services. Report submitted electronically

Director of CRI Bioinformatics Core: Jorge Atienza* Ph.D.

ABSTRACT

The goal of this study is to look for biomarkers that indicate the chances of patients with Ewing Sarcoma into prognostic groups and enable early detection of the biology of the disease Information analysis and machine learning approaches were applied on 36 Ewing Sarcoma patients along with their demographics. A German data set is provided as a validation set. A variety of statistical methods and computational experiments were used to identify gene signatures that can classify the patients into different survival states with high sensitivity and specificity.

1 PROJECT GOAL

The purpose of this study is to identify potential biomarkers that can segregate the Ewing Sarcoma patients into prognostic groups (e.g., normal status) and predict clinical outcomes. The main objective of this study is to develop a robust diagnostic tool for clinical decision making.

2 METHODS AND MATERIALS

2.1 Gene expression analysis of Ewing sarcoma patients

The gene expression data for Ewing sarcoma patients was obtained from the Broad Institute’s Gene Expression Atlas. The data set consists of 36 patients, with 18 in the training set and 18 in the validation set. The gene expression data was normalized and log-transformed. The differential expression analysis was performed using the DESeq2 package.

2.2 Identification of potential prognostic biomarkers

The potential biomarkers were identified using a two-step process. First, a univariate analysis was performed to identify genes that were significantly associated with survival. Then, a multivariate analysis was performed to identify a set of genes that were independently associated with survival.

3 RESULTS AND DISCUSSION

The results of the gene expression analysis revealed that 10 genes were significantly associated with survival. These genes were found to be differentially expressed between the two survival states. The multivariate analysis confirmed the reliability of these results. The identified genes were further validated using an independent data set.

4 CONCLUSION

The results of this study demonstrate the potential of using gene expression analysis to identify prognostic biomarkers for Ewing sarcoma patients. The identified genes can be used to develop a robust diagnostic tool for clinical decision making.
We also selected 5, 10, and 20 DE genes according to their VDB and calculated the expression profiles. The DE genes are compared with the random signature genes. Candidate signature genes based on VDB criteria can be used to reduce the number of redundant sampling of the signatures.

Positive gene signatures identified using random forest classifier in the CGG data set

By applying the procedure described in Section 2 and illustrated in Fig. 4, we selected and validated 169 candidate signatures from 35 DE genes for each size 5, 10, and 20 gene set. Table 1 shows the selected signatures with higher AUC among 399 signatures.

Table 1. Selected candidate gene signatures for the prediction of the survival status in the CGG data set

<table>
<thead>
<tr>
<th>Signature</th>
<th>AUC</th>
<th>Name</th>
<th>Gene</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGG00052</td>
<td>0.65</td>
<td>DE01</td>
<td>DE01</td>
<td>0.004</td>
</tr>
<tr>
<td>CGG00052</td>
<td>0.65</td>
<td>DE02</td>
<td>DE02</td>
<td>0.006</td>
</tr>
<tr>
<td>CGG00052</td>
<td>0.65</td>
<td>DE03</td>
<td>DE03</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Validation on the CGG data set showed good performance of the positive gene signatures from the CGG data set

An ideal prognostic gene signature derived from the CGG data set is expected to predict the survival status of German data set with high accuracy given the expression data from two data sets show similar distribution. However, we did not observe high predictive accuracy, which may be an explanation of the validation set using the best RF classification models from the CGG data set (data not shown).

We don't have enough information to answer the questions asked in the image. The text seems to be discussing the selection and validation of gene signatures for predicting survival status, but without more context or specific details, it's difficult to provide a more detailed response.
Established Workflows
RNA-seq
Exome

Documentation

Workshop Videos

CRI Galaxy Workflows

The following workflows were generated by bioinformaticians at the CRI and can be imported by each user and run with their data. Contact us if you have any questions. Additional workflows are being generated and will be added here as they are developed.

- RNA-Seq: Sample level (Paired-end) RNA-sequence analysis for individual samples which includes quality control, mapping and statistics for paired-end Illumina reads.
- RNA-Seq: Project level merge RNA-sequence analysis for merging multiple expressed list for single and paired-end Illumina reads.
- Exome Sequencing Analysis (Paired-End): for quality control, mapping, read statistics, and expressed list.
- Exome Sequencing Analysis (Single-Ended): for quality control, mapping, and expressed list.
Bioinformatics Core Training

- Introduction to R Programming
- Introduction to Linux Command Line for Bioinformatics
- Analyzing Illumina ChIP-seq Data with the CRI
- Introduction to CRI’s HPC Cluster for Bioinformatics Computing
- Analysis of Microarrays with R and BioConductor
- Analyzing Illumina Whole Exome Data with the CRI
- Galaxy: Web-based Bioinformatics Analysis & RNA-seq Workflow Management
The CRDW contains over >7 years of clinical data.

Research Enabling Tools

- Complete: 28
- In Progress: 6
- On Hold: 6

>225 completed requests

Patients:
- 607,000 encounters
- 6.1 million medications
- 16.1 million procedures
- 36.8 million labs
- 93 million diagnoses
- 13.9 million diagnoses
Data Requests

- Request submitted
- IRB approval or exemption
  - Data available?
  - Extract sample data
  - Meet PI for sign-off of sample data
  - Extract data
  - Deliver data to PI and receive sign-off
  - Request fulfilled

Office Hours
Tuesday 9-12
KCBD 1101
Center for Research Informatics

Bioinformatics

Clinical Research Data Warehouse

Applications Development

IT Operations and Infrastructure

Forms building
Data capture
Project workflow assistance
Dashboarding
Reporting
Applications support
REDCap support
Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN)
REDCap

Secure form and survey generation and data collection
Center for Research Informatics

- Bioinformatics
- Clinical Research Data Warehouse
- Applications Development
- IT Operations and Infrastructure

Group storage
Backup
Hosted Servers (VMs)
High-performance computing
Kenwood Data Center
Computational Resources

- **Storage**
  - 1.1 Pb of Isilon storage

- **Compute**
  - 2048 cores (2.2 GHz AMD Opteron 6274)

- **Self Service Data Analysis**
  - VMWare farm / up to 1500 VMs

- **Backup and Restore**
  - Spectralogic T950 / 300Tb of encrypted tapes

- **Large Memory Supercomputer**
  - Three - Intel® Xeon® E7-8870 @ 2.4GHz 1Tb RAM
FISMA-certified
HIPAA secure storage

Storage options are available.
More Information
http://cri.uchicago.edu

CRI Annual Report
http://crireport.uchicago.edu

General Questions
cri@bsd.uchicago.edu
Thank you.