Integrated biomedical data analysis utilizing various types of data for biomarkers identification

Ping Zhang  
*Menzies Health Institute*  
Griffith University  
QLD Australia  
p.zhang@griffith.edu.au

Amanda Cox  
*Menzies Health Institute*  
Griffith University  
QLD Australia  
a.cox@griffith.edu.au

Allan Cripps  
*School of Medicine*  
Griffith University  
QLD Australia  
allan.cripps@griffith.edu.au

Nicholas West  
*Menzies Health Institute*  
Griffith University  
QLD Australia  
n.west@griffith.edu.au

Abstract—Biomarkers discovery research requires the integrated analyses of a variety of the data across multiple domains, including clinical data, pathology data, gene expression, epigenetic data. Proper analysis can help understand the biological mechanism and better interpret the impact of the markers to disease. Realising the nature of the data in biomedical research and translational biomedicine, we developed a data analysis pipeline with a set of computational functions and an integrated method that can serve as a template for many biomarkers discovery research. The data analysis pipeline was developed with the data collected to identify biomarkers associated with obesity related disease. The set of functions included in the analysis template were used for finding the biomarkers and their combinatorial effect associated with obesity. The functions were developed in the general way that can be extended to other study easily.

Keywords—biomarkers, obesity, bioinformatics, biostatistics, high dimension, data analysis

I. INTRODUCTION

As technology grows, it is getting easier for medical researchers to collect large amounts of different types of data. A key issue we face now is maximizing the value of the data and optimizing the yield of research. Biomedical Big Data is becoming an emerging concept for discovery of biomarkers that help classify patients for personalized treatments. For biomarker discovery research, data collected typically include demographic and lifestyle data, clinical measures, images and various biological data. Luo et al [1] reviewed big data application in four major biomedical sub-disciplines bioinformatics, clinical informatics, imaging informatics, and public health informatics. One of the features of big data is variety of data types and structures [2, 3]. In biomedical research, very often, we collect a dataset with high dimensional variables, which is in big variety of data types, but in small samples size due to some cost issues. An example we have is the obesity study in our lab.

Many big data analytics tools have been developed and are already being used in biomedical research. As reviewed by Luo et al [1], the tools developed for different sub-disciplines may focus on different aspects, for example the tools in bioinformatics often need error identification and correction, but clinical informatics pays more attention to data-sharing, data security and need more specific ontologies for different issues. However, with our experience we realized that much biomedical research aiming for biomarker discovery needs particular methods from both bioinformatics and clinical informatics. Machine learning is often the main technique applied in these methods for building data analysis and prediction tools that help for the research. In a clinical setting, these tools can be used for looking at the molecular or sequencing data that predicts a drug response, whether lifestyle factors are associated with high risk of developing a particular chronic disease and whether blood tests can be used for early diagnosis of a disease. However, due to high dimension and small sample size of the data due to budgetary or feasibility issues, we need to use the tools very carefully. On the other side, different types of data need to be analysed in different types of analysis methods.

With obesity study in our lab as an example, we can see that it can be common to collect many types of data for biomarkers study. While factors such as lifestyle and caloric intake are major factors to body mass change, contributors to weight change, it is clear that there is a wide variation in individuals’ ability to maintain a healthy weight and metabolism, and in responses to specific dietary and exercise programs. Many key inflammatory markers, such as C-reactive protein, erythrocyte sedimentation rate, plasminogen-activator inhibitor 9 and cytokines, are associated with obesity [4]. Gut microbiota are increasingly recognized as having crucial roles in host functioning and may have impact on obesity [5]. GWAS have identified ~100 loci where polymorphisms show significant association with risk of obesity [6]. While estimates of the role of heredity in this variation are in the range of 40-70% [7], large scale GWAS can only attribute about 20% to common genetic variants [6]. A hypothesis supported by animal studies and increasingly by studies of human cohorts, is that epigenetic programming and its role as a mediator of interactions with the environment may be an important component of individual differences. On top of recognizing the dependency of microbiota profiles on life style, the impact of diet dependent microbial metabolites on the development and function of the immune system was addressed in [8]. The interaction between the factors can be complicated and may vary between individuals [9, 10]. It needs to be better
addressed and considered for intervention strategies to be developed. Despite some associations between obesity and single factors have been identified, the risks of disease have not been fully understood. To investigate various risk factors for obesity from different angles, our lab collected various types of data such as clinical measures, diet, immune gene expression data, 16S rRNA gene sequencing data and methylation sequencing data. Proper analysis can help better understand the complex biological network, interactions between different types of biomarkers and the impact to diseases.

The aim of this research is to develop a data analysis pipeline with a set of computational tools or functions and an integrated method for biomarkers discovery, which is suitable for the nature of the data in the research area of translational biomedicine. The data analysis pipeline can serve as a template for different biomarker studies that require analysis of various types of data. The set of tools helps integrate knowledge from different perspective for prediction of diseases. In this study we use obesity as an example of a complex disease to build the data analysis template due to the multifactorial nature of risk encompassing environmental, lifestyle, genetic, microbial and other biological factors. The main objective of the research is to identify the risk factors and the combinatorial effect of the risk factors associated with obesity, and to explore the interactions of the factors, for example how the environmental and epigenetic issues interact with immune functions or human gut microbiota change. Similar data analysis approaches would be suitable for biomarker discovery studies on other diseases, for example diabetes, Alzheimer’s and cancers. The functions were developed in a general way that can be extended easily to other datasets.

II. DATA ANALYSIS MODULES AND METHODOLOGY

The data used for this research included demographic data, clinical chemistry measures, flow cytometry data, Nanostring (nCounter Analysis System, https://www.nanostring.com) immune gene expression, 16S rRNA gene sequencing data from faecal samples and methylation sequencing data from gut mucosal samples. The analysis pipeline was developed in the way that each set of data can be analysed separately, and based on the analysis result from each data set, a feature selection method can be applied for selection of a combination of the biomarkers from the whole data set or from a selected subset of the data. Figure 1 is a diagram that depicts the integrated method we used for this study. Summarized descriptive statistics can be performed directly for demographic, clinical and flow cytometry data. Either mean and standard deviation, or median and range were calculated as the output depending on the distribution of the variable for continuous variables. Frequency or proportion of categories were calculated for categorical variables. Data cleaning and normalization of the data was performed separately before further analysis was completed through the pipelined functions. Based on the distribution of the data, t-test or Mann-Whitney test can be optional to be used for comparison of the continuous variables between 2 groups. For the categorical variables, Chi-squared test was developed in analysis pipeline to compare the variation between the groups. Principal component analysis (PCA) and hierarchical clustering analysis were developed as functions for exploratory analysis on selected data sets. Prediction models can be applied at each step of the analysis with different sets of data. Before further analysis can be applied, faecal sequences and methylation data can be processed by the functions that integrate published R packages into the pipeline. These include sequence alignment for identification of operational taxonomic units (OTUs), differentially methylated regions and associated genes.

A. Functions for Exploratory analysis

Descriptive statistics were implemented as one function that can be used for any data set that have been processed to be represented as a two-dimensional table, such as demographic, clinical and flow/mass cytometry data. The input for the function is a table that includes all the numeric or categorical columns with column (optional) to specify which diagnosis group that corresponding given individual belongs to. Output

Fig 1. Diagram of the data analysis pipeline. Different types of dataset can be read and analysed with a set of functions. The basic steps for each data set include, data cleaning, normalization, statistical tests and prediction models. For genetic data with different format, preprocessing steps can be different. During the analysis, database and knowledge management are involved.
for numerical variables include the mean, standard deviation, skewness, kurtosis, quantiles and normality test result. For categorical variables, proportion of each category is calculated. The output is in the form of a table. Inference Statistics functions including t-test and Mann-Whitney test, Chi-squared test for each variable were implemented to take .csv files as input and test result as output. For exploratory analysis, correlation analysis, hierarchical clustering and PCA functions were also implemented based on use of a two-dimensional table (data frame in R) as input.

For gene expression data, a set of normalization methods can be applied before the analysis. Standard comparison function between groups was designed and implemented with limma package and a volcano plot can be the output of the analysis based on the comparison result.

B. Analysis of faecal sequences and methylation data

In this study, faecal samples were sequenced with 16S-based approach. The pipeline for the analysis includes identification of OTUs, taxonomy classification, OTU abundance analysis, microbiota diversity analysis. Several bioinformatics pipelines are available for analysis of 16S rRNA data. Plummer et al [11] compared three commonly used ones, MetaGenome Rapid Annotation using Subsystem Technology (MG-RAST) [12], Quantitative Insights into Microbial Ecology (QIIME) [13] and mothur [14]. The limitation of the comparison was that authors only used default parameters of each of the pipeline for comparison of the results on taxonomic classification, diversity analysis and usability. In more recent bioinformatics studies focused on 16S sequencing data analysis, R package dada2 (DADA2) [15] was reported producing more accurate result than others. In this study, we focused on optimizing the parameters used by DADA2. The function designed for this analysis is to create the OTU table for further statistical and biological functional analysis.

Genome-wide DNA methylation profiles generated with the Illumina Infinium Human Methylation 450K BeadChip array (Illumina) was used as an example to develop part of the analysis pipeline. The Illumina 450K BeadChip assay allows for quantification of DNA methylation levels at 485,512 CpG dinucleotides. For each CpG site, a beta value was estimated, ranging from zero, representing completely unmethylated, to one, representing completely methylated sites. The function for identification of differentially methylated regions used limma [16] and DMRcate [17] packages from R. Based on the result, associated genes and biological pathways can be further analysed.

C. Analysis of cell type specific and pathway specific genes

Gene expression related to the certain cell type or a specific functional pathway may help reveal the biological mechanism of diseases. nSolver is a software for analysing nCounter data from NanoString technology. The advanced analysis from this software can include analysis of cell type specific genes. It can be used to compare the total abundance of the genes from a specific cell type between sample groups. This has a drawback when regulation of some of the genes in one cell type between sample groups is differential with regard to direction, that is, some genes are upregulated and others are down regulated potentially resulting in a zero net effect. In fact, the cell type or functional specific gene can be analysed with various modelling techniques. In our analysis pipeline, we developed a function that uses the cell specific and pathway specific genes [18, 19] to measure the abundance of the cell types. We also developed an analysis function to build a regression model for each of cell type and each functional pathway with identified corresponding genes. Due to the small sample size, which can be an often case, we developed the function with options of restricting the number of maximum genes to be included in the regression models, using forward stepwise selection or based on the gene expression fold change and p values from the exploratory analysis.

D. Predictive modelling and feature selection

Predictive modelling functions are developed based on different methods including linear regression, logistic regression, neural networks and support vector machines respectively. The functions were developed to read a two-dimensional table and output the cross-validated prediction result. Each column of the input table represents a clinical measurement or a gene (for gene expression data), and each row of the table represents an individual sample, for example a patient. Feature selection functions using stepwise or genetic algorithm developed and can be optionally called by any regression models. These functions are designed to be developed in separate modules in the pipeline, so they can be utilized in different stages of the analysis.

E. Integrated analysis pipeline

Analysis functions were developed in a way to be shared by each set of the data, including normalization functions, statistical tests, classification models and feature selection functions. Molecular sequencing data can be processed before applying these functions for analysis with different biological research questions. Genetic data can be analysed with separate methodologies. Sets of genes, cell types and functional networks or pathways can be selected and scored based on the outcome of the statistical analysis or the respective classification models. The scored genetic features can be combined with other selected variables for building a final prediction model.

III. EXPERIMENTAL RESULTS

This research used data from an investigation of the biomolecular factors in obesity as an example to test the method with the aim to find a potential set of obesity biomarkers from different types of data sets. The data included anthropometric data, clinical chemistry measures, flow cytometry data and immune gene expression for 28 individuals including 8 with BMI<25 and 20 with BMI>25 individuals.

A. Exploratory analysis result

Demographic, clinical and PBMC immune cell flow cytometry data were explored in descriptive analysis, hypothesis test and clustering analysis etc. A box plot file was an example output to show the comparison between normal weight and obese group for each of the variables included in
the dataset. Fig. 2 shows part of the output file. Fig. 3 is an example of the pairwise scatter plots which were created using the data from flow cytometric cell phenotyping to explore the correlation between each cell markers by group. Fig. 4 is a volcano plot produced from the genetic data for exploratory analysis. The result shows the differentially expressed genes (Fold change > 2 or Pvalue < 0.05 for this data set) between the two BMI groups. Fig. 5 shows the clustering analysis result in the form of heatmap, with the flow cytometric cell phenotyping data from paired peripheral blood mononuclear cells (PBMC).

B. Analysis of faecal sequences and methylation data

The faecal sequencing and methylation data were pre-processed using the functions that call the available R packages. OTU and taxonomic classification tables were produced before comparison analysis was performed with the exploratory analysis functions. For the obesity study, the microbial diversity were estimated and compared between two BMI groups. The richness estimators (Observed), nonparametric richness estimator (Chao1) and abundance-based coverage estimator (ACE) were used to calculate the observed richness of the microbial community. Shannon’s diversity index, Fisher’s alpha and Simpson’s probability were calculated for this study. The result can be merged with other measures for further analysis, see Table I. For more about the microbial community diversity metrics see [20]. There was little difference in the species richness or diversity between the groups (eg. Shannon index BMI low and high groups 3.78 ± 0.4 and 3.66 ± 0.43; p=0.42). PCA and hierarchical clustering analysis also showed none significant differences of taxonomic diversity between the two groups.

Table I. Microbial diversity analysis result

<table>
<thead>
<tr>
<th>Observed</th>
<th>Chao1</th>
<th>ACE</th>
<th>Shannon</th>
<th>Simpson</th>
<th>Fisher</th>
<th>sex</th>
<th>age</th>
<th>BMI</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>195</td>
<td>195</td>
<td>195</td>
<td>3.98</td>
<td>0.964</td>
<td>24.5</td>
<td>25</td>
<td>....</td>
<td>20.6</td>
<td>1</td>
</tr>
<tr>
<td>193</td>
<td>193</td>
<td>193</td>
<td>3.47</td>
<td>0.931</td>
<td>24.7M</td>
<td>54</td>
<td>....</td>
<td>38.7</td>
<td>2</td>
</tr>
<tr>
<td>162</td>
<td>162</td>
<td>162</td>
<td>3.44</td>
<td>0.928</td>
<td>18.9F</td>
<td>52</td>
<td>....</td>
<td>30.1</td>
<td>2</td>
</tr>
<tr>
<td>257</td>
<td>257</td>
<td>257</td>
<td>4.07</td>
<td>0.951</td>
<td>32.3</td>
<td>47</td>
<td>....</td>
<td>23.5</td>
<td>1</td>
</tr>
<tr>
<td>....</td>
<td>...</td>
<td>...</td>
<td>....</td>
<td>....</td>
<td>....</td>
<td>....</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>176</td>
<td>176</td>
<td>176</td>
<td>3.61</td>
<td>0.942</td>
<td>21.9F</td>
<td>36</td>
<td>....</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>180</td>
<td>180</td>
<td>180</td>
<td>3.92</td>
<td>0.953</td>
<td>22.0F</td>
<td>54</td>
<td>....</td>
<td>35.7</td>
<td>2</td>
</tr>
<tr>
<td>258</td>
<td>258</td>
<td>258</td>
<td>3.89</td>
<td>0.946</td>
<td>33.4F</td>
<td>30</td>
<td>....</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>271</td>
<td>271</td>
<td>271</td>
<td>4.15</td>
<td>0.972</td>
<td>33.6F</td>
<td>60</td>
<td>....</td>
<td>22</td>
<td>1</td>
</tr>
</tbody>
</table>
Similarly, the methylation levels of the differentially methylated genomic regions that were identified based on certain statistical confidence levels can also be merged to other data sets for further building complex prediction models.

C. Analysis of cell specific and pathway specific genes

Immune cell types, immune pathways and genes included in the NanoString nCounter® PanCancer Immune Profiling Panel (NanoString Technologies, Inc., Seattle, WA) were used for our experiments. This immune profiling panel includes 109 genes that define 24 immune cell types and 45 immune functional pathways. The average expression of the genes that are involved in each particular cell type are compared between two BMI groups, see Fig. 6.

To explore the contribution of each gene in each cell type or functional pathway, gene expression in each cell type or pathway were visualized separately. Fig. 7 shows each gene in T helper cells, as an example, compared between the two BMI groups. As we can see, between the two groups, the genes in the cells are not always up regulated for one BMI group. The logistic regression models we developed with the genes involved in each cell type and each functional pathway indicated that the genes involved in each type of cells as a combination do not distinguish the two BMI groups very well (highest validated classification rate 67% and AUC 0.778). The results from some models with genes included in immune functional pathways are shown in Table II. In this experiment, we restricted the maximum gene numbers to be included in the model to five due to the small sample size of our data. Some of the pathways, such as Th1 orientation pathway, regulation of inflammatory response and NK cell function were considered interesting and worthy for further investigation.

IV. CONCLUSION AND DISCUSSION

One aspect of Big Data is about data diversity. This research developed a template for helping biomedical research that involves a large variety of data types. Each subtype of the data can be processed separately and an overall profile of each individual sample can be merged from all the types of biological measures available. The set of analysis functions utilized bioinformatics and biostatistics techniques to analyse...
the multi-dimensional dataset. The translational approach to molecular research helps better understand the mechanisms underpinning the biological issues. In this study, the data set collected for obesity study was used as an example for testing the analysis methods. With the small sample size of the data, no strong biological conclusion has been drawn, though some preliminary result can suggest further investigations. Such as, the genes involved in inflammatory response pathway and Th1 orientation pathway have raised our interest for further investigation of their association with obesity. The methodology of the analysis pipeline was built for broad application that can be helpful for other complex diseases study. In some situations, some types of data are easy to be collected than other types for a same project. Therefore, often for a same project, more data may be available for one category of the data than another. For example, for the obesity study, it is likely to collect clinical data for more individual samples than genetic sequencing data due to the cost issues. In this case the pipeline can also be used with selected set of data or a whole set of data based on interests of the researchers or users. On the other side, the modular design of the pipeline allows new functional modules to be added at any time when needed. To create and visualize the correlation networks between selected variables from each types of data set is our next step to extend the functions of the analysis pipeline. At this stage, the pipeline is developed in R. It can be of use to many scientists with limited programming experience or to those with more experience who want to use this as a template to add in their own analysis functions.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>CV.err</th>
<th>AUC</th>
<th>Genes Involved</th>
<th>Total Genes in Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complement pathway</td>
<td>35.7%</td>
<td>0.865</td>
<td>CCL25, C1S, C1QB, C8G</td>
<td>9</td>
</tr>
<tr>
<td>Cytotoxicity</td>
<td>32.1%</td>
<td>0.797</td>
<td>PRF1, GZM4</td>
<td>10</td>
</tr>
<tr>
<td>Defense response to virus</td>
<td>32.1%</td>
<td>0.833</td>
<td>OAS3, PRF1, CXCL10</td>
<td>8</td>
</tr>
<tr>
<td>NK cell functions</td>
<td>25.0%</td>
<td>0.870</td>
<td>KLRF1, KLRG1, KLRD1</td>
<td>9</td>
</tr>
<tr>
<td>Regulation of inflammatory response</td>
<td>17.9%</td>
<td>1.000</td>
<td>CXCR2, CXCL9, CXCR1</td>
<td>16</td>
</tr>
<tr>
<td>Regulators of T cell activation</td>
<td>39.3%</td>
<td>0.802</td>
<td>TNFSF14, CD7, CD47</td>
<td>8</td>
</tr>
<tr>
<td>Regulators of Th1 and Th2 development</td>
<td>28.6%</td>
<td>0.786</td>
<td>CD7, CD5</td>
<td>2</td>
</tr>
<tr>
<td>Senescence pathway</td>
<td>35.7%</td>
<td>0.839</td>
<td>Hras, PLAU, CDKN1A</td>
<td>6</td>
</tr>
<tr>
<td>Th1 orientation</td>
<td>17.9%</td>
<td>1.000</td>
<td>IL18RA, IL18R1, IL1F1, IFNGR1</td>
<td>13</td>
</tr>
<tr>
<td>Toll like receptor</td>
<td>25.0%</td>
<td>0.979</td>
<td>TLR6, TLR5, TLR10, TLR7, TLR2</td>
<td>11</td>
</tr>
</tbody>
</table>

V. ACKNOWLEDGEMENT
This work is support by the Griffith University Area for Strategic Investment.

VI. REFERENCES


