Gene Expression Analysis of Acute Kidney Injury in Kidney Transplants

Kapil Panda

March 26, 2022

Abstract

In the world of kidneys, transplants are an option for patients in chronic stages that can save their lives, simultaneously can cause lots of additional issues and one of the most common effects being Acute Kidney Injury (AKI). Due to how rarely kidneys are biopsied, we are not able to find out much about what exactly is causing AKI, specifically in relation to gene expressions.

In this research, I have taken a dataset of 47 patients and their gene expressions: 28 with AKI, 8 who have had a complete nephrectomy, and 11 who are healthy patients. Using this data, I have trained a model using logistic regression that attempts to predict which patients are likely to have AKI depending on their gene expression levels. Using the model, I have been able to predict the top genes that are associated with having AKI and using a gene ontology analysis tool, I was able to find skeletal muscle terms that are associated with those genes.

1 Introduction

Every year, there are around 13.3 million cases of Acute kidney injury (AKI) which cause roughly 1.7 million deaths are caused by AKI. In the United States, AKI alone is associated with an increase in hospitalization costs that range from $5.4 to $24.0 billion. [usr] Even though there are such staggering amounts of cases each year, there still hasn't been concrete research done so far that concludes where AKI stems from, specifically from a genetic and molecular standpoint.

AKI is a condition of kidney failure or kidney damage that happens within a few hours or a few days where a build-up of waste products in the bloodstream occurs, making it hard for the kidneys to keep the right balance of fluid in the body. It is usually caused by a traumatic event that leads to kidney malfunction, such as dehydration, blood loss from major surgery or injury, or the use of medicines which cause decreased blood flow. The first symptoms of AKI typically include urinary tract blockage, body swelling, fatigue, or nausea.

Although AKI is an important problem, kidneys that have undergone AKI are rarely biopsied, which limits how much information we can get out of them.

However, kidney transplants offer a unique opportunity to study the injury-repair response to AKI because all kidney transplants experience some form of AKI. As part of the transplant process, numerous indication biopsies are performed to exclude rejection, guided by the international Banff histopathology consensus system. [XS;] Given that transplants are already an invasive procedure, biopsies can be collected with minimal additional risk or effort. Moreover, transplants have detailed function assessments and are followed indefinitely.

In this study, we examine a dataset centered around patients that experience AKI after kidney transplants. [ame] This dataset consists of a cohort of 47 individuals which underwent a kidney transplant, including 28 patients who had AKI, 11 patients who did not experience severe AKI, and 8 patients which underwent an entire nephrectomy. For each individual, the dataset measures the transcript expression levels for over 54,000 unique transcripts.

Throughout this research, most of the data revolved around gene expressions. Gene expressions are quite simply the instructions in our DNA that are converted into a functional product and help make important molecules such as proteins. All phenotypes stem from the expression of genes into transcripts, and then the translation of these transcripts into proteins. A lot of information can be found out about what a cell, organ, or tissue is doing, just by finding what genes are expressed. Being able to understand how gene expression is modulated in certain disease states can help us start to discover new ways to treat or prevent these diseases. Here, we examine the genes whose increased or decreased expression is associated with AKI in order to better understand its biology.

In this research, I constructed a logistic regression model to predict the identity of a cohort patient (i.e. AKI, normal, or nephrectomy) from the patient's set of normalized transcript expression levels. Logistic regression is a non-linear, machine learning architecture that predicts output probabilities for each possible AKI status. Like all regression analyses, logistic regression is used for training a predictive model and to describe data and explain the relationship between multiple variables. Logistic regression is a standard method of probabilistic regression, due to its simplicity in training (and therefore short training times), and its interpretability (i.e. we can examine the model and directly read off the input features that are driving a prediction).

2 Results

For each of the 47 patients in the cohort, I extracted their transcript expression levels over more than 54,000 transcripts. I then trained a logistic regression model to predict the probability each patient's AKI status (i.e. AKI, normal or nephrectomy). The input features for the model were the vector of transcript expression levels for each patient. Note that my model is trained to be a multi-class classifier, because it predicts between three classes, and each prediction is three probabilities of how likely it is that a patient belongs to each category.

On my training data of 47 patients, my model achieved 100% accuracy,

correctly classifying every single patient.

2.1 Table 1: Model accuracy for each class

Predictive class	Percent correctly classified
AKI	100%
Normal	100%
Nephrectomy	100%

Given the great performance of my model, I extracted the model weights for the AKI predictive class. This gives me a vector of over 54,000 multiplicative weights (one value for each transcript), which denote the signed importance the model gives to each transcript for predicting AKI.

2.2 Figure 1: Histogram of model weights for AKI

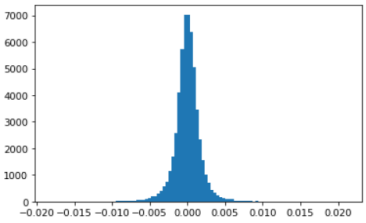

Figure 1: Histogram of weight values for the AKI predictive class. Positive values represent the genes that were high in patients with AKI as opposed to normal patients, and negative values represent genes that were high in normal patients compared to those with AKI.

Transcripts given very positive weights are associated with increasing the probability of AKI (according to the model); transcripts given very negative weights are associated with decreasing the probability of AKI. Most weights are centered around zero, which implies the model has learned that most transcripts are not informative for the prediction of AKI.

I translated the input feature transcripts into gene IDs, and considered the genes corresponding to the top 1000 most negatively predictive transcripts for AKI, and the genes corresponding to the top 1000 most positive predictive transcripts for AKI. I submitted these ranked genes to gene ontology enrichment using the GO Process ontology. [ENS+09]

2.3 Figure 2: GO Process enrichment for top negatively predictive transcripts

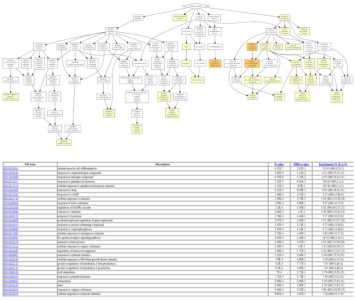

Figure 2: The GO Process enrichment tool shows which terms are most prevalent in the genes that were high in normal patients compared to AKI patients. The colored terms represent ones that are more significant. Results are also shown in a table in the second panel.

2.4 Figure 3: GO Process enrichment for top positively predictive transcripts

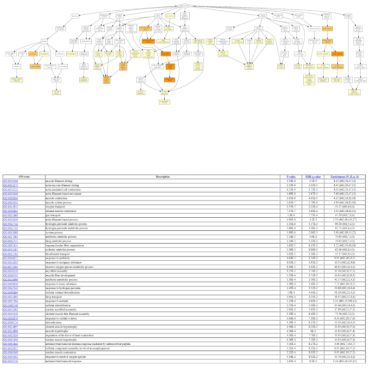

Figure 3: The GO Process enrichment tool shows which terms are most prevalent in the genes that were high in AKI patients compared to normal patients. The colored terms represent ones that are more significant. Results are also shown in a table in the second panel.

The negatively associated genes did not identify a strong pattern of associated GO terms, but the positively associated genes found a relatively strong association of skeletal-muscle-related terms with an increase of AKI probability (Figures 2-3).

3 Discussion

In this work, I trained a logistic regression model to predict AKI status from transcript expression levels. My model was high-performing, and then I examined the genes which were most positively and negatively predictive of AKI.

In my analysis, I found that skeletal muscle terms, including contraction processes (e.g. actin sliding) and muscle differentiation, were strongly associated with an increase in AKI. Although skeletal muscles may not be immediately obvious as being related to kidney injury, several previous studies have found associations between kidney damage and skeletal muscle processes. For example, rhabdomyolysis can be a cause of kidney damage, as toxins are released from decaying muscles. An existing association has already been found between AKI and kidney transplant donors with rhabdomyolysis. [F;]

It is important to point out the limitations of this association between skeletal muscle terms and AKI. Since the dataset and analyses are purely associative, we cannot conclude any form of causation between skeletal muscle terms and AKI. For example, rhabdomyolysis may be causing AKI, AKI may cause skeletal muscle genes to be expressed, or a third hidden factor may be causing both. After all, most machine learning is purely correlative.

Additionally, my dataset was fairly limited in size, with only 47 individuals. Although this gave me some potential associations, the small dataset may not be entirely representative of larger populations. Furthermore, I did not train with a validation or test set, as I treated the entire dataset as the training set. This implies that although my model may overfit (i.e. it may not generalize to new patients, despite its high performance on the training set), my analyses on what it's learned on this specific dataset remain valid. My predictive model may be used for future classification of patients, although because I did not have an available test set, its expected performance is not yet known.

In order to validate the results, further experimentation and larger datasets would be required. Further studies may rely on larger cohorts with more individuals with AKI, and future experiments (e.g. gene knockout experiments) may validate or refute the identified association of AKI with skeletal muscle physiology.

4 Methods

I downloaded processed data from BioGPS (Supplementary Information). I extracted transcript expression levels over 54,675 transcripts, for each of the 47 individuals, thus giving me a 47 x 54,675 design matrix.

Before training, I normalized the design matrix by mean-centering and variance-normalizing each feature (i.e. expression level vector) such that each transcript had a mean of 0 and variance of 1 across all patients.

I then trained a logistic regression model using SciKitLearn, to predict the AKI status (i.e. AKI, normal, or nephrectomy) from transcript expressions. Logistic regression takes the dot product of the transcript expression levels with a

learned weight vector, and adds a bias. For multivariate logistic regression (I am predicting three outputs not just one), there is one learned weight vector (and one bias) for each class, and the three output predictions are passed through a softmax for the final probabilities (the softmax is a multivariate extension of the sigmoid function). I trained this model with all the default parameters in SciKitLearn, except with a maximum of 1000 iterations.

Then, in order to extract the genes associated with the top negatively or positively predictive transcripts for AKI, I considered the learned weight vector just for the AKI output. I ranked the weights by the value, and took the transcripts with the top 1000 most negative weights, and the top 1000 most positive weights. I translated each transcript ID into a gene symbol using the AffyMetrix-published translation table. For transcripts that did not have a matched gene symbol, those transcripts were ignored. For transcripts that matched multiple gene symbols, I kept all matches.

I then inputed these top genes into GOrilla5, using the default parameters and reported the gene ontology enrichment results for the GO Process ontology.

5 Supplementary Information

Link to dataset: http://biogps.org/dataset/E-GEOD-30718/

Link to code: https://colab.research.google.com/drive/1rcfK4SI559oUJE9bRnv$_q$5$CzbSvX-UDsscrollTo = gJB31UrBjy7n$

References

[ame] Famulski, k. s., freitas, d. g. de, kreepala, c., chang, j., sellares, j., sis, b., einecke, g., mengel, m., reeve, j., halloran, p. f. (2012, may 1). molecular phenotypes of acute kidney injury in kidney transplants. american society of nephrology. retrieved september 11, 2021, from https://jasn.asnjournals.org/content/23/5/948.

[ENS+09] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. "gorilla: A tool for discovery and visualization of enriched go terms in ranked gene lists", from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-48, Feb 2009.

[F;] Lima RS;da Silva Junior GB;Liborio AB;Daher Ede F;. Acute kidney injury due to rhabdomyolysis. saudi journal of kidney diseases and transplantation : an official publication of the saudi center for organ transplantation, saudi arabia. retrieved september 17, 2021, from https://pubmed.ncbi.nlm.nih.gov/18711286/.

[int20] Acute kidney injury. international society of nephrology. (2020, august 25). retrieved september 11, 2021, from

https://www.theisn.org/commitment-to-kidney-health/focus-areas/acute-kidney-injury/., Aug 2020.

[usr] Annual data report. usrds. (n.d.). retrieved september 11, 2021, from https://adr.usrds.org/2020/chronic-kidney-disease/5-acute-kidney-injury.

[XS;] Chen CB; Zheng YT; Zhou J; Han M; Wang XP; Yuan XP; Wang CX; He XS;. Kidney transplantation from donors with rhabdomyolysis and acute renal failure. clinical transplantation. retrieved september 17, 2021, from https://pubmed.ncbi.nlm.nih.gov/28564273/.