Identifying Metabolic Signatures for Chronic Kidney Disease in Type II Diabetic Patients

Minya Pu[1], Rintaro Saito[2,3], Youyi Zhang[4]
Yurong Guo [2,3], Kumar Sharma[2,3], Loki Natarajan[1]
[1]Moores Cancer Center, University of California, San Diego, La Jolla, CA 92093-0901
[2]Insitute for Metabolomic Medicine, [3]Center for Renal Translational Medicine, University of California, San Diego, La Jolla, CA 92093-0711
[4]UTHealth, Graduate School of Biomedical Science, MD Anderson Cancer Center, Niles, TX 77030

Abstract
Diabetic patients with chronic kidney disease (CKD) are at much higher risk of morbidity, so it is of great importance to understand the disease mechanism. We used metabolomics data to explore the association between estimated glomerular filtration rate (GFR) values, a kidney function outcome, and a panel of urine biomarkers consisting of about 100 metabolites. A total of 114 type II diabetic patients were included in this analysis. We used two approaches, LASSO and k-TSP, to classify patients into DM+CKD and DM-CKD groups. We also used LASSO to explore the metabolites that were associated with disease severity in which eGFR values were used as a continuous outcome. A bootstrap-permutation based stability analysis was performed to assess the reproducibility of each variable in a LASSO model. We also showed that LASSO produced a lower leave-one-out cross-validation error rate than k-TSP in the training data set (1.3% vs. 6.6%), and also a slightly lower prediction error rate in the validation set (5.3% vs. 7.9%). A newer top scoring method may help to improve the error rates.

Key Words:
Classification, LASSO, Top scoring pair, Stability analysis, Metabolomics

1. Introduction

It is estimated that over 19 million adults and children in the US (8% of the population) have diabetes (ADA: http://www.diabetes.org/diabetes-basics/diabetes-statistics/). Diabetes can lead to serious complications, such as chronic kidney disease (CKD). Diabetes is the leading cause of kidney failure, accounting for 44% of new cases in 2011; over 200,000 people with end-stage kidney disease due to diabetes were living on chronic dialysis or with a kidney transplant in the United States in 2011 (ADA). Effective therapies to prevent diabetic patients from progressing to CKD have remained elusive. Protein and genetic biomarkers can provide insight into biological underpinnings of CKD within the context of existing diabetes but may have limited value in a primary metabolic/environmental based disease such as diabetic complications. Metabolomics is a systematic evaluation of small molecules and may provide fundamental biochemical insights into disease pathways. Urine metabolomics offers a wide range of measurable metabolites (Sweetman, *et al*, 1971; Nyhan, *et al*, 1969; Aramaki, *et al*, 1989) as the kidney is responsible for concentrating a variety of metabolites and excreting them in the urine. In addition, urine metabolomics may offer direct insights into biochemical pathways linked to kidney dysfunction.

We have previously identified a signature of diabetic kidney disease that separated patients with diabetes and CKD from healthy controls (Sharma, *et al*, 2013). This led to

novel insights into the underlying biochemical basis for diabetic kidney disease. However, from a clinical perspective, it will be more useful to identify a subset of metabolites that best discriminates diabetics with CKD (DM + CKD) from diabetics without CKD (DM – CKD). By separating out the effects of diabetes itself from the markers of diabetic kidney disease, we will gain a better understanding on which patients are at greatest risk for complications. This will provide more insight into markers of renal dysfunction, and subsequently elucidate possible therapeutic targets for treating CKD. Thus, we aimed to develop a multivariate metabolomic signature that discriminates DM+CKD from DM-CKD, using the established definition of CKD as estimated glomerular filtration rate (eGFR) < 60 ml/min/1.73m^2. In addition, to examine predictors of the full range of kidney function, we also identified metabolic biomarkers that predicted the continuous spectrum of eGFR values. In this analysis we focused on type II diabetes (T2DM), the more common diabetic condition, accounting for 90% of adult diabetes. In addition, evaluating markers of CKD amongst T2DMs, rather than including all diabetic types (i.e. Type I and Type II) reduces the chance of confounding since disease etiology and biological signatures are likely to differ between Type I and Type II diabetics. Data from a cohort of 114 diabetic patients were utilized.

A large number of statistical methods have been developed to classify patients into two or more classes, such as penalized logistic regression (Zhu & Hastie 2004), decision trees (Breiman et al, 1984) and random forests (Breiman 2001), nearest neighbor methods (Ripley 1996), linear discriminant analysis (McLachlan 2004), bagging and boosting (Breiman 1996; Alfaro et al, 2013), support vector machines (Bennett & Campbell 2002), prediction analysis of microarrays (PAM) (Tibshirani, *et al*, 2002) and many others (Ripley 1996; Venables & Ripley 2002; Lu & Han 2003). Here we focused on two recent statistical approaches LASSO (Tibshirani, 1996; Friedman 2008) and k-TSP (Geman *et al*, 2004; Tan, *et al*, 2005; Leek, 2009; Damond, 2011). Both are state-of-arts approaches that have been shown to be efficient for variable selection when there are a large number of predictors, so that one can build parsimonious models for classification. LASSO is a regression based approach and k-TSP is rank-based. Detailed description about these two methods can be found in Section 3.

In this paper, in Section 2, we described the study sample and the data processing steps for the metabolomics data. In Section 3 and 4, we briefly introduced the two statistical methods LASSO and k-TSP, and then we developed and compared classifiers using these two methods; a rigorous training versus validation paradigm was used. We also proposed a novel approach to perform stability analysis for models selected by LASSO; using this approach, a selection p-value was calculated for each selected variable in the models to assess the variable significance. In section 5, we discussed our contribution of the stability analysis to the LASSO method, summarized our findings and introduced new developments in the top scoring approach; at the end, we described our plans for future approaches.

2. Data collection and processing

2.1 Materials, methods and patient samples

As previously noted, we used eGFR values as a main measure for kidney function and grouped patients into the DM+CKD group if eGFR <60 ml/min/1.73m^2 at the time of 24h urine collection and in the DM-CKD group otherwise. eGFR was determined from two serum creatinine measurements based on the four-variable Modification of Diet in Renal

Disease study equation (Levey, 1999). Gas chromatography-mass spectrometry was employed to quantify 104 urine metabolites from a total of 114 type II diabetic patients. These patients were participants from our previous studies; they came from different geographic regions from the United States and Finland (Groop, *et al*, 2009; Sharma, *et al*, 2011; Sharma, *et al*. 2013). Stratified by CKD status, the sample was randomly split into a training set (N= 76) for model development and validation set (N=38) for testing the model on an independent set.

Baseline Characteristics of the DM+CKD vs DM-CKD patients were compared using Kruskal-Wallis tests or Fisher's Exact tests (Table 1). On average, compared to the DM-CKD patients, the DM+CKD patients were marginally older, more likely to be non-white, and had higher BMI, lower blood pressure, longer diabetic duration, lower HbA1c %, and higher serum creatinine level. Age, gender, race, BMI, DM2 duration, MAP (mean arterial pressure = (2xDiastolic BP+Systolic BP)/3), HbA1c %, and urine ACR were included in the analyses below as predictors for eGFR outcomes.

Table 1. Baseline characteristics (mean +/- SD for continuous variables, count (%) for binary variables) of the study sample.

	1:training DM+CKD	2:training DM-CKD	3:validation DM+CKD	4:validation DM-CKD	p
n	49	27	24	14	
Age, years	64.7 +/- 9.9	59.4 +/- 7	63.1 +/- 8.5	58.5 +/- 6.7	0.014
Race: non-white	25 (52%)	0 (0%)	10 (42%)	1 (7%)	4.40E-07
Race: white	23 (48%)	27 (100%)	14 (58%)	13 (93%)	
Gender Male Female	31 (63%) 18 (37%)	14 (52%) 13 (48%)	17 (71%) 7 (29%)	9 (64%) 5 (36%)	0.59
BMI, kg/m2	34.7 +/- 7.3	23.9 +/- 2.8	31.1 +/- 5.6	26.5 +/- 4.7	5.50E-11
Smoking: Never	22 (45%)	10 (37%)	10 (48%)	7 (50%)	0.84
Smoking: Ever	27 (55%)	17 (63%)	11 (52%)	7 (50%)	
Systolic BP, mmHg	131 +/- 14.8	136.6 +/- 16	131.2 +/- 17.5	143.4 +/- 14.7	0.037
Diastolic BP, mmHg	70.2 +/- 7.2	80.6 +/- 8.7	73.1 +/- 10.4	86.2 +/- 7.1	4.00E-08
Type 2 DM Duration, years*	16 (10, 25)	11 (8, 18)	13 (9, 18)	10 (8, 14)	0.034
HbA1c, %	7.3 +/- 1.2	8.6 +/- 1.3	7.3 +/- 1.2	8.1 +/- 1.3	0.00084
Serum Creatinine, mg/dL	2.2 +/- 0.7	0.9 +/- 0.1	2.1 +/- 0.8	1 +/- 0.2	1.90E-16
Albumin/Creatinine Ratio*	0.19 (0.04, 1.22)	0.08 (0.04, 0.13)	0.26 (0.14, 0.82)	0.14 (0.06, 0.35)	0.11
eGFR, ml/min/1.73m^2	35.2 +/- 11.1	82.8 +/- 16.6	35.9 +/- 11.6	74 +/- 10.2	6.30E-17

*Median and interquartiles are presented.

2.2 **Data filtering, manipulation and imputation**

A total of 104 metabolites were considered. Metabolite distributions were examined and assessed for excessive number of zeros and missing values, and 87.9% of the metabolites

did not have any missing values. Among those with missing values, the proportion of missing values per metabolite ranged from 1% to 47.5%. Five metabolites were excluded from the analysis set due to a large number of missing values (> 15%). If over one-third of values for a metabolite were zero, that metabolite was dichotomized to 1 if it had a nonzero value, and 0 otherwise; 5 were excluded due to fewer than 15% nonzero values. The remaining 99 metabolites (71 continuous, and 28 dichotomized) were used for analysis. Occasional (often < 7.5%) missing values were imputed using the mean for that variable across all subjects with available data. If a binary metabolite variable had missing values, its value was imputed as a random draw from a Bernoulli distribution with proportion parameter set equal to the proportion of nonzero values in the observed non-missing data for that metabolite. A log2 transformation was applied to the 71 continuous variables; if a metabolite value was observed to be zero, a quantity of half of the minimum value of that metabolite was added to ensure that all values were non-zero before applying the log-transformation.

3. Statistical methods

We used a split-sample training-validation paradigm to avoid overfitting, and to obtain optimism-corrected estimates of prediction error. We randomly allocated patients into training and validation sets based on a ratio of 2:1 and stratified on CKD status, so we had 76 patients in the training set and 38 in the validation set. All models were developed using only the training set; prediction accuracy was evaluated and the different classifiers were compared using the validation set.

3.1 eGFR as a binary outcome

We first developed models to discriminate between diabetics with and without CKD (DM+CKD or DM-CKD) based on the clinical cutpoint of eGFR <60 ml/min/1.73m^2.

3.1.1 Univariate analysis.

A univariate Welch's t-test was first used to compare each continuous metabolite marker (log2 transformed) between T2DM patients with and without CKD in the training set. For dichotomized variables, a Fisher's exact test was used to compare the two groups. P-values were adjusted to control false discovery rate (FDR), using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Metabolites with FDR$\leqslant$ 0.05 in this training set analysis were then compared in the validation set samples, and metabolites with Bonferroni (Abdi, 2007) adjusted p-value $\leqslant$ 0.05 were considered to be validated, and deemed to be significantly different between DM+CKD and DM-CKD groups.

3.1.2 Multivariate analysis

The univariate approach, useful for identifying differentially expressed metabolites, does not provide an algorithm for classifying metabolites to disease groups. Multivariate approaches are needed to develop classifiers. We first used Principal components analysis (PCA) to reduce dimensionality and examine if the first few principal components are able to separate the DM+CKD vs DM-CKD groups (data not shown). A disadvantage of PCA is that all metabolites are used to create the principal components, so that a subset of most predictive metabolites is not selected. Thus, to identify parsimonious metabolite sets, we applied two methods, the LASSO and k-TSP, to elicit a subset of metabolites

that could be used to discriminate the DM+CKD versus DM-CKD classes. We briefly describe these methods below.

3.1.2.1 LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is a powerful tool for model selection when the number of the predictors is large (Tibshirani, 1996; Park and Hastie, 2007). Different from generalized linear models, it adds a L_1-regularized term, which results in variables with strong associations with outcomes being selected, while those with weak associations omitted because their coefficients are set to zero. Since our outcome is a binary variable reflecting CKD status, the LASSO model solves for β by minimizing the following objective function:

$$|\log(p/(1-p)) - X\beta|^2 + \lambda|\beta|_1$$

where p is the probability of having CKD, $|.|^2$ is the usual residual sum of squares, and $|v|_1$ is the L_1-norm of a vector v. Five-fold cross validation and a grid search was used to determine λ_{min}, the L_1-regularization parameter that minimized the misclassification error. However, as recommended, we used the value $\lambda min+1se$ for the parameter to avoid overfitting (Waldmann 2013), where $\lambda min+1se$ is the largest λ such that the error is within 1 standard error of the minimal mis-classification error rate.

Before running LASSO, all the variables were standardized. The analysis was implemented using the R-glmnet package (Friedman, *et al*, 2010).

3.1.2.2 k-Top scoring pairs (k-TSP)

The Top Scoring Pair classifier is a rank-based approach that searches initially for a pair of features (e.g., genes, metabolites) whose expression values switch most consistently (Geman, *et al*, 2004; Xu, *et al*, 2005; Leek, 2009). It classifies patients into disease groups based on the relative expression of that pair, say X_i and X_j. The score was defined as the absolute difference of two marginal probabilities $Pr(X_i<X_j|DM+CKD)$ and $Pr(X_i < X_j |DM-CKD)$. The top scoring pair would be the one that achieves the largest score. For prediction, suppose $X_i<X_j$ classifies patients into DM+CKD in the training set, for a new sample, if $X_{i,new} < X_{j,new}$, then the "new" sample is classified as belonging to DM+CKD; otherwise, the sample is classified to DM-CKD. If there are tied TS pairs, a secondary score is created to pick the pair that has the largest average rank difference between two classes (Tan, *et al* 2005). To improve performance, k disjoint top pairs can be used together for classification (Tan, *et al* 2005; Damond, 2011), and a majority voting procedure is used to determine the predicted class of a new patient. For this reason, k is set to be odd numbers only, i.e. k=1, 3, 5, 7, 9, etc. Cross validation is used to determine the optimal choice for k to minimize classification error rate. It has been shown that TSP/k-TSP is simple and accurate, and it is very easy to interpret the results. Furthermore, it often outperforms some complex machine learning methods such as k-nearest neighbor and naïve Bayes (Tan, *et al* 2005).

We applied the TSP/k-TSP algorithm to identify pairs of metabolites for classifying patients into DM+CKD vs DM-CKD groups. All the continuous predictors were standardized so that all the variables were comparable. This analysis was performed using the R-ktspair package.

3.2 Metabolites and eGFR values on a continuous range

To examine metabolite predictors of the entire range of kidney function, we used eGFR values as a continuous outcome.

3.2.1 Univariate analysis:

Spearman rank correlation was used to explore the association between a continuous predictor and the outcome, and a Wilcoxon rank sum test was used for a binary predictor. Other steps were the same as those used in Section 3.1.1.

3.2.2 Multivariate analysis

A linear model with variable selection implemented via the LASSO was fitted to the training set. Five-fold cross validation was used to select the regularization parameter that minimized mean absolute error. Predictors included the 99 metabolites, and all the clinical variables mentioned in section 2.1 right above Table 1. The outcome eGFR values were log2 transformed. Root mean square error (RMSE) and Spearman correlations were calculated on the validation data set for the observed and predicted eGFR values from the final LASSO model. For this analysis, the TSP approach is not applicable as it only applies to categorical outcomes.

3.3 Bootstrap-permutation based stability analysis for LASSO

We examined if the selected metabolites were reproducible, i.e., would we obtain the same "predictive" set of metabolites for a different sample of diabetics with and without CKD drawn from the same underlying population as the study sample? Therefore, we conducted the following stability analysis to evaluate reproducibility of the models.

To evaluate the stability of a selected model, first, we drew N bootstrap samples of the same size of the training set from the training data set. LASSO was run on each of these bootstrap samples. The regularization parameter was re-estimated for each bootstrap sample. The proportion of the times that each metabolite was chosen was calculated. To be more specific, suppose a variable *x* was selected M times via the LASSO procedure in N bootstrap samples, this proportion is P=M/N. Metabolites chosen with high probability across the bootstrap samples were regarded as stable.

Second, we used a permutation test approach to assess if the proportion of a variable being chosen, P=M/N, was "significantly" higher than by mere chance. To assess this, outcome labels were randomly permuted (DM+CKD vs. DM-CKD for the binary outcome or individual eGFR values for the continuous outcome) K times. For the k*th* permuted sample, N bootstrap samples were obtained and LASSO was re-run to obtain P_k ($=M_k/N$), where M_k is the number of times the variable *x* was chosen in the k*th* sample; k=1, 2, …, K.

Lastly, we assigned a selection p-value to each variable in the selected model by calculating Prob ($P_k > P$), where k goes from 1 to K.

In this paper, we used N=500 and K=1000.

3.4 Model comparisons

For the binary outcome, i.e. DM+CKD vs. DM-CKD, models selected by LASSO and k-TSP were compared by computing (1) leave-one-out cross validation mis-classification error rate using the training data set, and (2) prediction error rates using the validation data set. In calculating (1), in the LASSO model, the regularization parameter was re-estimated at every round; in k-TSP, we fixed k=3 every round, but three pairs of variables could differ each time.

4. Results

4.1 Metabolite signature of DM+CKD and DM-CKD groups

4.1.1 Univariate analysis comparing DM+CKD and DM-CKD groups

Univariate t-tests or Fisher's exact tests showed that among 99 metabolites, from the training set, 75 were significantly different between DM+CKD and DM-CKD patients after FDR adjustment ((FDR < 0.05). Based on the validation set, 45 of these were still found to be significant after Bonferroni adjustment. The significant results are shown in Supplementary Table 1.

4.1.2 Multivariate classification of DM+CKD and DM-CKD groups

LASSO for variable selection: The LASSO method selected 10 variables to be included in the final model (Table 2). Direction of the association with CKD was indicated by the signs of the coefficients (i.e., positive coefficients indicate increased risk of CKD). Re-fitting the models based on 500 bootstrapped samples showed that the first 6 variables were quite robust because over 60% of the models selected them. Stability analysis showed that, 8 out of these 10 variables had a p-value<0.05. Exclusion of these two insignificant variables should have little impact on the prediction performance.

Table 2: Model selected by LASSO using eGFR as a binary outcome (DM+CKD vs. DM-CKD) with stability analysis results. Variables were sorted based on selection probabilities.

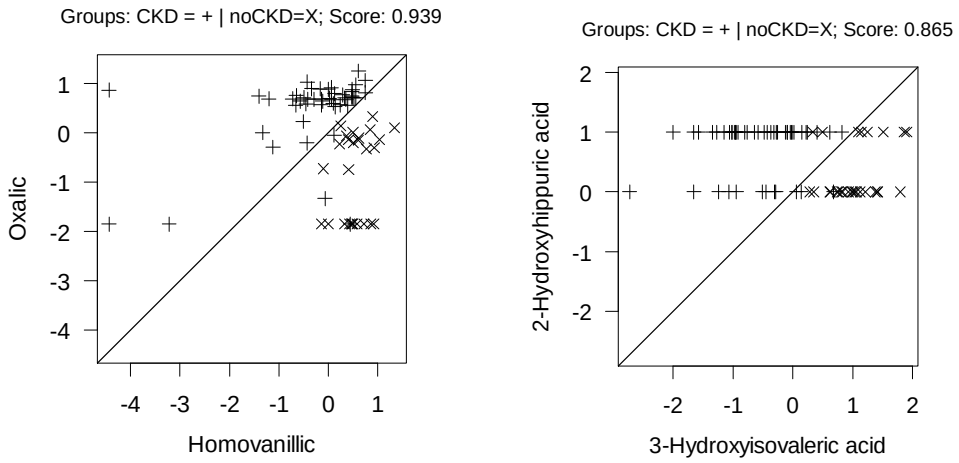
Variable (HMDB number for metabolites)	Mean coefficients	P Probability of being selected	Prob. (P>P_k)
Glycolic Acid (00115)	-0.5775	0.886	0.001
Sebacic Acid, binary (00792)	-0.3372	0.786	0.002
3-Hydroxy Isovaleric Acid (00754)	-0.3729	0.744	0.001
Methylmalonic acid, binary (00202)	-0.2055	0.642	0.005
BMI	0.1644	0.606	0.033
Oxalic Acid (02329)	0.2180	0.604	0.021
2-Hydroxyadipic acid (00321)	-0.0910	0.382	0.035
Aconitic Acid (00072)	-0.0753	0.354	0.033
4-Hydroxybutyric acid, binary (00710)	-0.050	0.222	0.091
4-Hydroxyisovaleric acid, binary (02011)	-0.0001	0.210	0.093

Classification using k-TSP

When TSP was applied, i.e., only one pair of the predictors was used to classify patients, two metabolites, homovanillic acid and oxalic acid, were chosen as the highest scoring pair. When using k-TSP, two more pairs of metabolites were chosen. Five-fold cross validation from the k-TSP method chose k=3. The three pairs of metabolites were used together to classify patients: a patient was classified to the DM+CKD group if the following two or more criteria were satisfied: Homovanillic acid < Oxalic acid; 4-Hydroxyisovaleric acid (binary) <Succinylacetone (binary); 3-Hydroxyisovaleric acid

< 2-Hydroxyhippuric acid (binary) (Fig. 1). Since the second pair had both dichotomized variables, they were not plotted. Note that the continuous variables were standardized.

Figure 1. Two of the three pairs of metabolites selected by k-TSP to predict CKD vs. no CKD. The second pair 4-Hydroxyisovaleric acid and Succinylacetone-1-4 were both binary variables was thus not plotted. All the continuous predictors were standardized.



4.1.3 Comparisons of the discrimination of LASSO and k-TSP

Based on the training set, TSP gave a cross-validation error rate of 0.079, which was slightly higher than the rate from using k-TSP which was 0.066 (Table 3). Using more variables, the model selected by LASSO had even lower leave-one-out cross-validated mis-classification error rates compared to k-TSP (0.013 vs. 0.066). LASSO also produced a lower prediction error rate than k-TSP (0.053 vs. 0.079) using the validation data.

Table 3 Classifiers chosen by LASSO, TSP and k-TSP.

method	LASSO	TSP	k-TSP
Classifier	10 variables as shown in Table 2	Homovanillic <Oxalic	Homovanillic <Oxalic; 4-Hydroxyisovaleric acid (binary) < Succinylacetone (binary); 3-Hydroxyisovaleric acid < 2-Hydroxyhippuric acid (binary)
Leave-one-out cross-validation error rate on training set	0.013	0.079	0.066
Prediction error rate on validation set	0.053	0.156	0.079

4.2 Metabolite predictors of the range of kidney function
4.2.1Univariate analysis using eGFR as a continuous outcome

Univariate Spearman rank correlation tests or Wilcoxon rank sum tests showed that among 99 metabolites, from the training set, 88 were significantly associated with eGFR levels after FDR adjustment (FDR < 0.05). Based on the validation set, 45 of these were still found to be significant after Bonferroni adjustment. These significant results can be found in supplementary table 2. Among them, 23 of were also found significant in the univariate analysis performed in section 4.1.1.

4.2.2 LASSO model using eGFR as a continuous outcome

LASSO chose 17 metabolites and 3 clinical variables (DM2 years of duration, MAP and age) as the best predictors of eGFR values (Table 4). A stability analysis of 500 bootstrapped samples showed that these variables were quite robust and stable. We noticed that the significance of most (13/17) metabolic variables was confirmed (i.e. selection p<0.05) by our bootstrap-permutation based method, the significance of the clinical variables (age, MAP, and DM2 duration) did not reproduce. This implies that the metabolic changes have greater or more immediate impact on CKD progression than a patient's clinical characteristics.

The Spearman rank correlation using the validated set between the observed and fitted eGFR values was 0.68, p<0.001, and the mean absolute error between these two sets of values in validation set was 0.36 (95%CI 0.23 -0.48).

Table 4: Model selected by LASSO using eGFR as a continuous outcome with stability analysis results. Variables were sorted based on selection probabilities.

Variable (HMDB number for metabolites)	Mean coefficients	P Probability of being selected	Prob. (P>P_k)
Oxalic acid (02329)	-0.0978	0.908	0.002
3-Hydroxyisovaleric acid (00754)	0.0888	0.826	0.001
Glycolic acid (00115)	0.1046	0.822	0.001
3-Hydroxybutyric acid, binary (00357)	0.0773	0.774	0.002
Oxoadipic acid, binary (00225)	-0.0372	0.738	0.041
Fumaric acid (00134)	-0.0374	0.686	0.02
Type 2 DM duration, years	-0.036	0.652	0.1
Mean Arterial Pressure (MAP)	0.0326	0.646	0.079
Palmitic acid (00220)	-0.0322	0.636	0.014
3-Hydroxydecanedioic acid, binary (00350)	0.0334	0.628	0.038
4-Hydroxyisovaleric acid, binary (02011)	7.00E-04	0.618	0.006
2-Hydroxyhippuric acid, binary (00840)	-0.0297	0.596	0.14
Citric Acid (00094)	0.0449	0.556	0.018
Age	-0.0245	0.556	0.2
3-Hydroxyadipic acid (00345)	0.0188	0.500	0.22
Methylmalonic acid, binary (00202)	0.0339	0.476	0.0499
Aconitic acid (00072)	0.0263	0.448	0.042
Uric acid (00289)	0.0152	0.372	0.14
4-Hydroxyphenylpyruvic acid, binary (00707)	-0.0117	0.358	0.33
Succinylacetone, binary (00635)	-0.0157	0.332	0.035

5. Discussion

Selecting a set of biomarkers that discriminate disease subgroups could improve our understanding of disease progression, and could eventually impact diagnostic and treatment approaches. When the set of biomarkers is large, identifying a "best" subset can be a computationally challenging and statistically difficult because of the potential to overfit and obtain spurious results. Traditional methods such as step wise logistic regression based on AIC would fail. Given the large number of predictors, by chance some of them will be able to make a perfect separation between response and nonresponses. In addition, a model obtained from a stepwise procedure is often very unstable and does not validate in independent samples (Breiman, 1992). Univariate analysis could provide a hint of which metabolites might be relevant. However, when there are a large number of predictors, a greater penalty is needed to correct for multiple testing, which will cause important but only weakly significant predictors to be filtered out before a subsequent analysis such as Gene Set Enrichment Analysis might be performed. Also, many of the metabolites selected as significant in the univariate analysis were redundant. Thus, we employed modern advanced statistical techniques to develop metabolomic signatures for disease classes.

Our main contribution is that we have proposed a bootstrap-permutation based method to assess the significance of each variable selected by LASSO, as described in section 3.3. We first run LASSO on bootstrapped samples to obtain the observed probability that each variable is selected; next, we permute the outcome labels in the original data set and use bootstrapping again on these permuted sets to generate a null distribution for these selection probabilities. The significance of these probabilities can be assessed by comparing the observed selection probabilities against their own null distributions. Our approach is straightforward, intuitive, and it was quite useful. As illustrated in an example in section 4.2.2 in which the outcome was a continuous variable, a large number of variables were chosen by LASSO, and some of them will surely drop out from the model when there were only slight changes in the data set, for reasons such as missing values on a few subjects. Thus, it was important to assess what variables were truly important or truly stable and what were in fact negligible among those that were selected.

We used a well-researched parametric method LASSO as well as a conceptually simple rank-based k-TSP approach to build predictive models. These are two effective methods, and LASSO seems to work slightly better in our study, as shown in Table 3 of Section 4.1.3. LASSO achieved a prediction error of 5.3% in the validation samples whereas the corresponding error rate for k-TSP was 7.9%. In spite of this difference, they produced non-overlapping predictive sets which could be both very useful in exploring the mechanism of the disease progression.

There have been new developments related to the top scoring method since it was first proposed in 2004. Among them, TST (top scoring triplets) was proposed to use three genes together for classification (Lin et al, 2009), and then there was a further extended version, TSN, where N ($N \geqslant 2$) genes were used together to generate scores (Magis & Price, 2012). Furthermore, instead of using difference in marginal probabilities to produce scores, the Chi-squared statistic was proposed and the corresponding method was called TSG (Wang, *et al*, 2013). Here we used the original TSP/k-TSP approach and

we will utilize and compare these newer methods in future work. Using a newer method such as TSG may improve the performance.

Future work will also include the interpretation of these selected models in terms of their biological relevance. We have developed a novel bioinformatic tool called rsMetabPPI that can relate these significantly differentiated metabolites to the metabolic pathways in diabetic kidney disease (Saito, 2014). rsMetabPPI integrates currently known human metabolic networks with publicly available human protein-protein interaction networks. We will also evaluate the metabolites chosen from these analyses and the statistical tools to larger datasets of patients with diabetes and the presence or absence of kidney disease. As there are many controversial definitions as to what constitutes kidney disease with diabetes, the application of metabolomics and advanced statistical tools to address this issue is of major clinical importance.

In summary, we applied a popular variable selection technique, the LASSO, to identify a metabolomics signature of CKD among diabetics. As a novel contribution, we proposed a bootstrap-permutation method to assess stability of the signature. Furthermore, we compared the more complex model-based LASSO approach to a simple non-parametric method, TSP/kTSP, using a rigorous training-validation paradigm for developing and evaluating the classification rule. In future work we aim to develop biological metabolite/protein interaction networks associated with the metabolomics signature with an eye towards improving diagnosis and treatment of CKD among diabetics.

Acknowledgements

Funding for this project was supported by a grant from National Institute of Diabetes and Digestive and Kidney Diseases (1DP3DK094352-KS) and a grant from National Cancer Institute/National Institute of Health (R01CA166293).

References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In Salkind, N. J. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

Alfaro, E., Gamez, M. and Garcia, N. (2013). adabag: An R Package for Classification with Boosting and Bagging. *Journal of Statistical Software*, Vol 54, 2, pp. 1–35.

Aramaki, S., Lehotay, D., Nyhan, W.L., MacLeod, P.M., and Sweetman, L. (1989). Methylcitrate in maternal urine during a pregnancy with a fetus affected with propionic acidaemia. *Journal of Inherited Metabolic Disease*, 12:86-88.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289–300.

Bennett, K.P., Campbell, C. (2002). Support vector machine: Hype or hallelujah? *SIGKDD Explorations*, 2(2).

Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth Statistics/Probability Series. Wadsworth Advanced Books ans Software, Belmont, CA.

Breiman, L. (1992). The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error. *Journal of American Statistical Association*, 87, 738-754.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, Vol 24, 2, pp.123–140.

Breiman, L. (2001). Random Forests, *Machine Learning*, 45(1), 5-32.

Damond, J. (2011). Presentation and study of robustness for several methods to classify individuals based on their gene expressions. Master thesis, Swiss Federal Institute of Technology Lausanne (Switzerland).

Friedman, J., Hastie, T., Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software*, Vol. 33(1), 1-22.

Geman, D, d'Avignon, C., Naiman, D. and Winslow, R. (2004). Classifying gene expression profiles from pairwise mRNA comparisons. *Statistical Applications in Genetics and Molecular Biology*, 3: Article 19.

Groop, P.H., Thomas, M.C., Moran, J.L., Waden, J., Thorn, L.M., Makinen, V.P., Rosengard-Barlund, M., Saraheimo, M., Hietala, K., Heikkila, O; FinnDiane Study Group. (2009). The presence and severity of chronic kidney disease predicts all-cause mortality in type 1 diabetes. *Diabetes*, 58:1651-1658.

Leek, J.T. (2009). The tspair package for finding top scoring pair classifiers in R. *Bioinformatics*, 25(9):1203-4.

Levey, A.S., Bosch, J.P., Lewis, J.B., Greene, T., Rogers, N., Roth, D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation. *Ann Intern Med* 130: 461– 470.

Lin, X., Afsari, B., Marchionni, L., Cope, L., Parmigiani, G., Naiman, D., Geman, D. (2009). The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations. *BMC Bioinformatics*, 10:256.

Lu, Y., Han, J. Cancer classification using gene expression data. *Journal Information Systems - Special issue: Data management in bioinformatics*, Volume 28 Issue 4, Pages 243 – 268.

Magis, A.T., Price, N.D. (2012). The top-scoring 'N' algorithm: a generalized relative expression classification method from small numbers of biomolecules. *BMC Bioinformatics*. 2012 Sep 11;13:227.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience. ISBN 0-471-69115-1.

Nyhan, W.L., James, J.A., Teberg, A.J., Sweetman, L., and Nelson, L.G. (1969). A new disorder of purine metabolism with behavioral manifestations. *J Pediatr* 74:20-27.

Park, M.Y., Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B*, 69:659– 677.

Ripley, B.D. (1996). Pattern Recognition and Neural Networks. Cambridge University Press.

Sharma, K., Ix, J.H., Mathew, A.V., Cho, M., Pflueger, A., Dunn, S.R., Francos, B., Sharma, S., Falkner, B., McGowan, T.A., et al. (2011). Pirfenidone for diabetic nephropathy. *J Am Soc Nephrol* 22:1144-1151.

Sharma, K., Karl, B., Mathew, A.V., Gangoiti, J.A., Wassel, C.L., Saito, R., Pu, M., Sharma, S., You, Y.H., Wang, L., Diamond-Stanic, M., Lindenmeyer, M.T., Forsblom, C., Wu, W., Ix, J,H,, Ideker, T., Kopp, J.B., Nigam, S.K., Cohen, C.D., Groop, P.H., Barshop, B.A., Natarajan, L., Nyhan, W.L., Naviaux, R.K. (2013). Metabolomics reveals signature of mitochondrial dysfunction in diabetic kidney disease. *J Am Soc Nephrol*, 24(11):1901-12.

Sweetman, L., and Nyhan, W.L. (1971). Detailed comparison of the urinary excretion of purines in a patient with the Lesch-Nyhan syndrome and a control subject. *Biochem Med* 4:121-134.

Tan,A.C., Naiman, D.Q. , Xu, L., Winslow, R.L. , Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21: 3896-3904.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* (Methodological), Vol. 58, No.1, 267-288.

Tibshirani,R. et al.(2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci.* USA, 99, 6567–6572.

Venables, W. N., Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth edition. Springer.

Waldmann, P., Mészáros, G., Gredler, B., Fuerst,C. and Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, 4:270.

Wang, H., Zhang, H., Dai, Z., Chen, M.S., Yuan, Z. (2013). TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection. *BMC Med Genomics*. 2013; 6 Suppl 1:S3.

Saito, R., Pu, M., Natarajan, L., Ideker, T.,Sharma, K. (2014). A Novel Bioinformatic Tool for Identifying Proteins Regulating Metabolites. *American Journal of Physiology – Renal Physiology*. To be submitted.

Xu, L., Tan, A.C., Naiman, D.Q., Geman, D., Winslow, R.L. (2005) Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, **21**, 3905–3911.

Zhu J., Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. Biostatistics, 5(3):427-43.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67(2):301–320.

Supplementary Table 1. Univariate results comparing DM+CKD and DM-CKD patients. Only Significant results were presented here.

	Variable	Raw p-value	FDR adjusted p-value	Bonferroni corrected p-value
1	B_GLUTARIC	7.00E-08	2.20E-07	1.70E-07
2	B_GLYCOLIC	6.10E-19	3.00E-17	1.00E-06
3	B_3.OHPROPIONIC	0.00011	0.00019	3.50E-06
4	B_CITRIC	2.40E-14	4.00E-13	2.30E-05
5	B_3.OHISOVALERIC	6.10E-19	3.00E-17	2.80E-05
6	B_2.ME.30H.BUTYRIC..1.2.	4.70E-15	1.60E-13	3.10E-05
7	B_AZELAIC	1.70E-10	1.20E-09	6.10E-05
8	B_HIPPURIC	2.60E-08	9.40E-08	8.80E-05
9	B_SEBACIC.di	3.00E-10	1.90E-09	2.00E-04
10	B_SUBERIC	3.30E-09	1.40E-08	0.00037
11	B_ISOVALERYLGLYCINE.1.2.	2.20E-09	1.10E-08	0.00059
12	B_2.ETHYL.3.OHPROPIONIC	6.30E-12	6.30E-11	0.00088
13	B_ETHYLMALONIC	2.40E-07	5.40E-07	9.00E-04
14	B_GLYCERIC	4.30E-10	2.50E-09	0.001
15	B_ACONITIC	2.40E-13	3.40E-12	0.0012
16	B_2.OHADIPIC	1.90E-14	3.70E-13	0.0013
17	B_ADIPIC	3.20E-05	5.90E-05	0.0017
18	B_3.HYDROXYISOBUTYRIC	7.90E-15	2.00E-13	0.0019
19	B_3.OHGLUTARIC	3.00E-09	1.40E-08	0.0019
20	B_METHYLCITRIC	4.80E-11	4.00E-10	0.0023
21	B_2.OHISOVALERIC	1.60E-08	6.00E-08	0.0026
22	B_2.OHPHENYLACETIC	1.10E-12	1.30E-11	0.003
23	B_4.OH.HIPPURIC	1.10E-07	2.80E-07	0.003
24	B_4.OHCYCLOHEXYLACETIC..1.2.	1.40E-06	3.00E-06	0.003
25	B_3.ME.ADIPIC	1.20E-12	1.30E-11	0.0061
26	B_INDOLELACTIC	2.80E-08	9.80E-08	0.0073
27	B_DECADIENEDIOIC..1.2..di	1.40E-07	3.30E-07	0.0083
28	B_4.OHBUTYRIC.di	9.20E-08	2.50E-07	0.012
29	B_4.OHISOVALERIC.di	9.20E-08	2.50E-07	0.012
30	B_HEXANOIC	4.70E-10	2.60E-09	0.012
31	B_3.MEGLUTARIC.di	0.00065	0.00094	0.013
32	B_5.HIAA	0.00011	0.00019	0.015
33	B_N.ACETYLTYROSINE.1.2.	5.30E-05	9.30E-05	0.016
34	B_PHENYLLACTIC	1.10E-07	2.80E-07	0.016
35	B_4.OHPHENYLLACTIC	2.20E-06	4.40E-06	0.018
36	B_OCTANOIC	9.90E-08	2.70E-07	0.018
37	B_ISOBUTYRYLGLYCINE..1.2..di	3.10E-05	5.80E-05	0.021
38	B_ISOCITRIC	8.60E-08	2.50E-07	0.021
39	B_3.OHBUTYRIC	7.60E-08	2.30E-07	0.022
40	B_3.MECROTONYLGLYCINE.1.2.	2.20E-10	1.50E-09	0.027
41	B_HOMOVANILLIC	5.90E-06	1.20E-05	0.029
42	B_OROTIC	2.00E-09	9.80E-09	0.035
43	B_MEVALONIC.di	0.00016	0.00026	0.036
44	B_METHYLSUCCINIC	1.40E-07	3.30E-07	0.043
45	B_URACIL	7.50E-12	6.80E-11	0.046

Supplemental Table 2. Univariate results assessing the association of each metabolite with continuous eGFR levels. Only Significant results were presented here.

	Variable	Raw p-value	FDR adjusted p-value	Bonferroni corrected p-value
1	B_2.OXOADIPIC..S.A..di	3.98E-29	7.97E-28	9.83E-14
2	B_PHENYLPROPIONYLGLYCINE..1.2..di	2.17E-29	7.97E-28	9.83E-14
3	B_2.MEBUTYRYLGLYCINE..1.2..di	2.96E-29	7.97E-28	1.40E-13
4	B_HEXANOYLGLYCINE..1.2..di	1.57E-29	7.97E-28	1.40E-13
5	B_PROPIONYLGLYCINE..1.2..di	3.98E-29	7.97E-28	2.63E-13
6	B_4.OHPHENYLPYRUVIC..S.A..di	1.87E-28	2.33E-27	4.50E-13
7	B_TRANSCINNAMOYLGLYCINE0.di	2.87E-28	2.61E-27	4.50E-13
8	B_2.OXOBUTYRIC..A.S..di	2.87E-28	2.61E-27	5.70E-13
9	B_GLUTACONIC.di	5.30E-29	8.83E-28	5.70E-13
10	B_ISOBUTYRYLGLYCINE..1.2..di	2.87E-28	2.61E-27	7.07E-13
11	B_LINOLEIC.di	5.08E-28	4.24E-27	7.07E-13
12	B_PALMITOLEIC.di	1.17E-28	1.67E-27	7.07E-13
13	B_SUCCINYLACETONE..1.4..di	8.27E-28	5.91E-27	1.03E-12
14	B_METHYLMALONIC.di	2.06E-27	1.09E-26	1.20E-12
15	B_MEVALONIC.di	7.10E-28	5.46E-27	1.38E-12
16	B_SUBERYLGLYCINE..1.2..di	1.24E-27	8.27E-27	1.56E-12
17	B_2.OH.HIPPURIC.di	2.52E-27	1.09E-26	1.73E-12
18	B_OLEIC.di	1.56E-27	9.73E-27	1.73E-12
19	B_SEBACIC.di	2.22E-27	1.09E-26	1.89E-12
20	B_3.OHVALERIC.di	2.77E-27	1.09E-26	2.15E-12
21	B_2.OXO.3MEVALERIC.S.MU.1620.di	2.87E-27	1.09E-26	2.24E-12
22	B_4.OHBUTYRIC.di	2.77E-27	1.09E-26	2.24E-12
23	B_4.OHISOVALERIC.di	2.77E-27	1.09E-26	2.24E-12
24	B_2.MEGLUTACONIC..E.Z..di	2.22E-27	1.09E-26	2.29E-12
25	B_3.MEGLUTARIC.di	3.06E-27	1.09E-26	2.29E-12
26	B_3.OHDECANEDIOIC.di	3.01E-27	1.09E-26	2.31E-12
27	B_DECADIENEDIOIC..1.2..di	3.01E-27	1.09E-26	2.31E-12
28	B_PHENYLACETIC.di	2.52E-27	1.09E-26	2.31E-12
29	B_GLYCOLIC	2.81E-15	9.38E-15	5.13E-08
30	B_3.OHISOVALERIC	2.28E-15	7.85E-15	3.34E-06
31	B_AZELAIC	2.63E-10	7.31E-10	2.97E-05
32	B_CITRIC	4.78E-11	1.37E-10	7.20E-05
33	B_2.ME.30H.BUTYRIC..1.2.	3.11E-12	9.72E-12	8.67E-05
34	B_3.HYDROXYISOBUTYRIC	9.44E-12	2.86E-11	0.0001067
35	B_ACONITIC	3.70E-14	1.19E-13	0.0001453
36	B_2.ETHYL.3.OHPROPIONIC	3.95E-10	1.07E-09	0.000419
37	B_3.OHPROPIONIC	0.0001724	0.0002573	0.0004672
38	B_2.OHISOVALERIC	9.34E-06	1.70E-05	0.00126
39	B_ISOCITRIC	3.73E-11	1.10E-10	0.001272
40	B_2.OHGLUTARIC	0.0002145	0.0003154	0.002735
41	B_HIPPURIC	7.01E-05	0.0001079	0.005046
42	B_GLUTARIC	0.0007523	0.001017	0.006833
43	B_URACIL	9.82E-10	2.52E-09	0.007077
44	B_SUBERIC	2.09E-05	3.55E-05	0.01794
45	B_OCTANOIC	1.48E-07	3.52E-07	0.04734