PUEPro: A Computational Pipeline for Prediction of Urine Excretory Proteins

Yan Wang[1], Wei Du[1, 4,*], Yanchun Liang[1], Xin Chen[1, 4],
Chi Zhang[4], Wei Pang[3], Ying Xu[1, 2, 4,*]

[1]College of Computer Science and Technology
[2]College of Public Health, Jilin University, Changchun, China
[3]School of Natural and Computing Sciences, University of Aberdeen,
Aberdeen, AB243UE, UK
[4]Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology
and Institute of Bioinformatics, University of Georgia, Athens, GA, USA
*Correspondence Authors: weidu@jlu.edu.cn; xyn@uga.edu

Abstract. A computational pipeline is developed to accurately predict urine excretory proteins and the possible origins of the proteins. The novel contributions of this study include: (i) a new method for predicting if a cellular protein is urine excretory based on unique features of proteins known to be urine excretory; and (ii) a novel method for identifying urinary proteins originating from the urinary system. By integrating these tools, our computational pipeline is capable of predicting the origin of a detected urinary protein, hence offering a novel tool for predicting potential biomarkers of a specific disease, which may have some of their proteins urine excreted. One application is presented for this prediction pipeline to demonstrate the effectiveness of its prediction. The pipeline and supplementary materials can be accessed at the following URL: http://csbl.bmb.uga.edu/PUEPro/.

Keywords: urine excretory proteins; support vector machine recursive feature elimination; biomarkers of disease

1 Introduction

Early detection is essential for disease control and possible prevention [1]. Among the existing techniques, detection of biomarkers in body fluids such as blood, urine or saliva represents the least invasive and most efficient approaches, which can offer an initial indication of diseases in specific organs. A key to accomplishing this lies in our ability to accurately identify informative biomarkers. Technical challenges involve (1) accurate identification of overly produced biomolecules in targeted disease tissues, which are specific to the disease, and (2) reliable prediction of which of such biomolecules can enter a specific type of body fluid.

Compared to blood, urine is probably equally information rich in term of the types of biomolecules from different origins. This makes urinary biomarkers more desirable, considering that (i) urine tends to have a simpler composition, which simplifies

adfa, p. 1, 2011.

the detection problem compared to blood; (ii) the dynamic range across different proteins is substantially smaller in urine than in blood; and (iii) collecting urine is substantially less invasive and easier to do than blood collection.

Proteins in urine originate mainly from glomerular filtration of serum proteins [2] and from the urinary system through secretion and membrane shedding. Therefore, it is necessary to identify and remove proteins that are from the urinary system among proteins found in urine, in order to identify biomarkers for diseases in distal organs [3]. Currently, the most useful disease markers in urine have been largely for urogenital diseases, such as urothelial cancer [4], renal cell carcinoma [5], prostate cancer [6], and bladder cancer [7]. A few recent studies have demonstrated the feasibility in using urinary proteins as disease markers in distal organs, such as ovarian carcinoma [8], lung cancer [9], hepatocellular carcinoma [10], and gastric cancer [11].

Only a few studies have been published on the prediction of urinary proteins, ours being one of them [11]. The present study extends the previous study by including novel capabilities for identification of origins of detected proteins in urine in addition to an improved prediction tool for proteins that are urine excretory. Our study utilizes a few data sources of urinary proteins to build a predictor for such proteins, including those given in [12-16]. The current knowledge is: 70% of the urinary proteins originate from the kidney and the urinary tracts, and the remaining 30% are filtered from blood circulation by the glomerulus [17]. Specifically, the origins of urinary proteins are: (i) glomerular filtration of blood proteins; (ii) proteins from renal tubular epithelial cells and other urinary cells, including those secreted from these cells or shed from their plasma membranes; (iii) membrane shedding proteins from renal tubular epithelial cells and other urinary system; (iv) exosome secretion; and (v) the whole cell shed from urinary tracts [2, 18].

A few studies have been published on the identification of the origins of detected urinary proteins, such as the work presented in [19], which identified urinary proteins which originated from kidney using an isolated rat kidney model, and studies that identified urinary proteins as being from the urinary tracts [3, 20]. These data are used to train our computational predictor for the origins of detected urinary proteins. Overall, the current study has made the following novel contributions: (i) a novel approach to predicting excretory proteins in urine; and (ii) a novel method for predicting the origins of detected urinary proteins. A server called PUEPro (Prediction of Urine Excretory Proteins) has been developed based on these novel methodological developments, and it can be accessed at: http://csbl.bmb.uga.edu/PUEPro/, from which the supplementary files of this paper can be downloaded.

2 Material and Methods

2.1 Data collection

Collecting urinary proteins and generating negative training data.

Several datasets of proteins have been identified in human urine, including those in the Sys-BodyFluid database [12] and the Human Proteome Project (HPP) database [13]. The Sys-BodyFluid database consists of 1,941 distinct human proteins that have

been experimentally identified in nine urinary proteomic studies. Over 2,000 experimentally verified urinary proteins are available and retrieved from the HPP database. In addition, we have also gathered urinary proteins identified by other urinary proteomic studies [14-16]. Overall, a total of 3,133 unique human urinary proteins were collected. To rule out the possibility of false identification of urinary proteins, we have used 1,495 out of the 3,133 proteins that have been detected by more than one study as the positive data in our study. Among the 1,495 proteins, 1,000 are used as training data and the remaining 495 as the test data.

Since we do not have a very clear understanding about which cellular proteins cannot be excreted to urine, generating a negative dataset is a challenge. In this study, we applied a selection process similar to the one presented by Cui et al. [21] through choosing proteins from the Pfam protein families [22] that do not contain any proteins that have been detected in urine. For each Pfam family (with at least ten members), ten members are randomly selected as part of the negative data. As a result, 1,821 proteins are selected as the negative data, of which 1,000 are used as training data and 821 proteins as the test data.

Collecting urological proteins in urine and generating negative training data.

A few studies have been published regarding identified urinary proteins with originate from the urinary system. For example, 990 human proteins are predicted to be homologs of rat kidney proteins [19]. In addition, other studies have identified more urinary proteins with origins in the urinary system [3, 20]. We have compared these proteins with the above 1,495 urinary proteins, and found that only 430 of them originates from the urinary system. To predict which urinary proteins do not originate from the urinary system, we used a similar procedure discussed earlier, i.e., to select proteins from the Pfam families which do not contain any of the 430 proteins detected above. This gives us 365 urinary proteins which do not originate from the urinary system.

2.2 Model construction

Feature construction.

We aim to identify sequence or structure-based features that can distinguish between a specified positive set and a negative set as discussed in the previous section. We have examined features of the following types: (1) general sequence features; (2) physicochemical properties; (3) specific domains/motifs; (4) structural properties. The general sequence features include sequence length, amino acid composition, autocorrelation and quasi-sequence-order of each protein. The physicochemical properties include hydrophobicity, polarity, charge, secondary structure, and molecular weight. Specific domains/motifs include transmembrane score, signal peptide, and the number of glycosylation sites. Structural properties include secondary structure composition, radius of gyration among a few others. Overall, 39 features, represented by 1,537 feature elements, are considered and are shown in Table S1 of the supplementary material.

Distinguishing feature selection.

For these features elements, there are four major categories: relevant features, redundant features, irrelevant features, and noisy features. For the feature set containing many features, the relevant features are only very small part of the whole feature set, and most of the features are irrelevant features. So, many feature selection methods for expression data analysis remove the irrelevant features firstly. In this research, we have employed a two-stage feature-selection procedure to distinguish the positive datasets from the negative ones on the training dataset. A t-test, which is a simple and effective filter feature selection method, was used to determine and eliminate the features without discerning power for our problem. Based on the calculated p-value, a q-value for each feature was calculated to control the False Discovery Rate [23]. Q-value = 0.005 was used as the threshold of q-value for removing non-contributing features. In the second step, a support vector machine (SVM)-recursive feature elimination (RFE) procedure [23], which is one of the best embedded feature selection methods, was applied to rank the remaining features, and to remove the lowly ranked and non-contributing features by the backward elimination technology, which selects relevant features by iteratively removing the most irrelevant feature at one time until the predefined size of the final features subset is reached. In each loop, the feature ranking of the remaining features can be possibly modified. At the end, 87 feature elements were selected and used in our analysis. The method eliminates irrelevant features and selects relevant features according to a criterion related to their support to a discrimination function *DJ*, which is measured by training SVM at each step. The discrimination function *DJ* is defined as follows:

$$\begin{cases} H = y_i y_j K(x_i, x_j) \\ DJ(i) = (1/2)a^T Ha - (1/2)a^T H(-i)a \end{cases}, \quad (1)$$

where y_i and y_j are the class labels of samples x_i and x_j. $K(x_i, x_j)$ is the kernel function that measures the similarity between x_i and x_j. α is obtained by training the clasifier of SVM in the algorithm of SVM-train. The algorithm of SVM-RFE [23] is defined as follows:

SVM-RFE Algorithm:

Input:

Training examples: $X_0 = [x_1, x_2, \ldots x_i, \ldots x_n]^T$

Class label: $y = [y_1, y_2, \ldots y_i, \ldots y_n]^T$

Initialize:

Subset of surviving features: $s = [1, 2, \ldots m]$

Feature ranked list: r=[]

Repeat until $s = []$

Restrict training examples to good feature indices: $X = X_0(:, s)$

Train the classifier by SVM : $\alpha = SVM\text{-}train(X, y)$

Compute the matrix H: $H = y_i y_j K(x_i, x_i)$

Compute the ranking criteria: $DJ(i) = (1/2)\alpha^T H\alpha - (1/2)\alpha^T H(-i)\alpha$
Find the feature with the smallest ranking criterion: f = *argmin*(*DJ*)
Update feature ranking list: r = [s(f), r]
Eliminate the feature with smallest ranking criterion: s = s(1:f-1, f+1:length(s))
End
Output:
Feature ranking list r.

Classification and assessment.

For determining the class labels for the new proteins correctly, a classifier needs to be constructed. In this research, the Support Vector Machine (SVM) is used as the classifier with several evaluation criteria, which are used to guide the choice of parameters. In SVM, the hyperplane of a high dimensional space, which is called feature space, is constructed to separate two classes. A good separation of one hyperplane needs to have the largest distance to the nearest training data of any class. The kernel functions, the wide coefficient of kernel functions, and the penalty coefficient *C* are the main parameters of SVM. Gaussian kernel with a single parameter *q* is a common choice for classification. Then, we can select the combination of *C* and *q* by grid search to improve the effectiveness. Using different parameters for the classifier, we can derive the distance *d* between the positions of the prediction data in the feature space and the optimal separating hyperplane. A larger distance *d* means more reliable prediction results. The SVM-based classifier was trained on the training data using the selected features to predict if a protein is urinary or not. Similarly, a second classifier is trained to predict whether a urinary protein originates from the urinary system.

The following measures are used to evaluate the prediction performance: $\text{specificity}=\frac{TN}{TN+FP}$, $\text{precision}=\frac{TP}{TP+FP}$, $\text{accuracy}=\frac{TP+TN}{N_{total}}$, $\text{MCC}=\frac{(TP\times TN-FP\times FN)}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}$ and AUC (the area under the curve) of the sensitivity-specificity curve [24], where TP is the number of true positives, FP refers to the number of false positives, TN means the number of true negatives, FN for the number of false negatives, and N_{total} is the total number of proteins for prediction in a given test set.

2.3 Identification of differentially expressed genes

We have applied our prediction method developed above to the gene-expression data of lung cancer versus control samples. The dataset consists of RNA-seq data of 101 paired samples of lung cancer and control samples, which are downloaded from the TCGA database [25]. The following formula is used to estimate the fold-change of each gene:

$$FC_{ij} = \begin{cases} fc_{ij} - 1 & (fc_{ij} \geq 1) \\ 1 - \dfrac{1}{fc_{ij}} & (fc_{ij} < 1) \end{cases}, \text{ where } fc_{ij} = C_{ij} \big/ N_{ij}, \quad (2)$$

$$FC_i = \frac{1}{m} \sum_{j=1}^{m} FC_{ij} \quad , \quad (3)$$

where m is the number of samples, and C_{ij} and N_{ij} are the expression levels of gene i in the jth pair of cancer and normal control. If FC_i is greater than zero, the relevant gene is considered as up-regulated in cancer; otherwise down-regulated or no change. 0.5 (-0.5) was used as the threshold for defining differentially expressed genes. In addition, a Wilcoxon test was used to assess the statistical significance of the observed differential expression in cancer vs. normal samples, and the statistical significance cutoff value is set at 0.05.

3 Results

3.1 Features of urine-excretory proteins

39 features, represented as a vector of 1,537 elements (see Table S1), are used to distinguish between the positive and the negative training data by training an SVM-based classifier with an RBF kernel. 10-fold cross-validation was done to assess the performance of the trained classifier, and the classifier has the average sensitivity, specificity, precision and accuracy of 84.40%, 82.87%, 83.20%, and 83.63%, respectively.

A feature selection process was then conducted to select the most discerning parameters among the 1,537. At the end, 87 parameters were selected, which achieve comparable performance to the above. Among the selected parameters, transmembrane domains and signal peptides have been found to be useful for predicting protein secretion to blood circulation [11, 21]. The radius of gyration is an interesting one, which has been suggested to play a role in proteins passing through the GBM (glomerular basement membrane). Published studies have observed that proteins with a radius smaller than 1.8 nm can pass through the GBM-slit diaphragm barrier, whereas proteins with a radius larger than 4.0 nm are retained [26].

3.2 Performance of urine-excretory proteins

Urine-excretory proteins prediction.

We then retrained an SVM-based classifier based on the 87 selected parameters, using both the linear and RBF kernel. The performance assessment of the classifier was done using 10-fold cross-validation by repeating the prediction 100 times on the training set containing 1,000 positive and 1,000 negative samples. It is found that the classification accuracy ranges from 81.00% to 97.00% for the positive data and from

74.51% to 94.12% for the negative data. The average performance based on the linear and RBF kernel is shown in Table 1. In addition, the ROC curve is given in Figure 1 (left).

Table 1. Average performance of urine protein prediction by 10-fold cross validation on training set

SVM Kernel	Sensitivity	Specificity	Precision	Accuracy	MCC	AUC
Linear	88.77%	88.70%	88.61%	88.74%	0.775	0.947
RBF	88.05%	87.57%	87.52%	87.81%	0.756	0.937

We then assessed the trained models on an independent test set composed of 495 urine-excretory proteins and 821 non-urine excretory proteins, with the detailed protein names given in Table S2 of the supplementary material. The prediction performance is presented in Table 2 along with the ROC curve in Figure 1(right). At the end, we have selected the classifier using the RBF kernel as it performs better than the linear model on the test set.

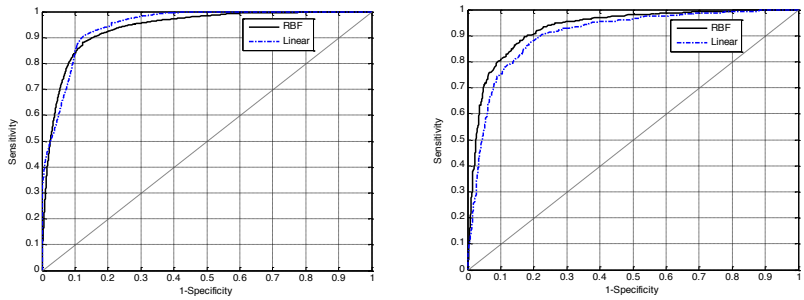

Fig. 1. The ROC curve for 10-fold cross validation on the training set (left) and on an independent testing set (right).

Table 2. Average **performance** of our classifier on the independent testing set

SVM Kernel	Sensitivity	Specificity	Precision	Accuracy	MCC	AUC
Linear	83.84%	83.19%	75.05%	83.43%	0.658	0.906
RBF	83.84%	87.82%	80.58%	86.32%	0.712	0.931

Predicting and ranking the known excretory proteins in urine.

We define the D-value of a protein in the UniProt database [27] as follows:

$$D = d * p \quad , \tag{4}$$

where p = 1 if the protein is predicted to be urine-excretory and -1 otherwise; and d is the distance between the position of the protein in the feature space and the separating hyper-plane defined by the trained SVM classifier. 228 (22.8%) of the positive training data (1,000) are ranked among the top 1,000 proteins. Among these 1,000 proteins, 110 (22.2%) are in the positive test dataset (495).

We also ranked the urine-excretory proteins that have been detected to be associated with human diseases in the literature and do not overlap with our training data. To accomplish this, we have collected such proteins from the Urinary Protein Biomarker Database [3]. 261 proteins are found to be in both this database and the UniProt, with the detailed protein names given in Table S3. 56 (21.46%) of these 261 proteins ranked among the top 1,000, 91 (34.87%) among the top 2,000 and 123 (47.13%) among the top 3,000, as detailed in Table 3 along with the p-values being 6.45e-21, 4.22e-28, and 1.83e-35, respectively. A comparison was also included in Table 3 between the results by our model and by a previous study [11], which is the only relevant study in the literature. In our model, we have employed a two-stage feature-selection procedure to distinguish the positive datasets from the negative ones on the training dataset and applied SVM with different kernel functions.

We have also conducted a function enrichment analysis of the top 1,000 D-value ranked proteins, using DAVID [28] against the Gene Ontology, KEGG, BBID and BIOCARTA databases, and using the whole set of UniProt as the background set. The goal is to check the subcellular locations as well as the biological processes enriched by these proteins. For understanding the cellular functions and subcellular locations of these predicted excretory proteins in urine, we noted that the most significantly enriched biological processes and cellular components were cell adhesion and extracellular region. In addition, the most significantly enriched pathways are cell adhesion molecules, ECM-receptor interaction, and complement and coagulation cascades (see Table S4), which are all closely involved in the urine excretory process.

Table 3. A comparison among the ranking results of known urinary biomarkers for diseases by our classifier *versus* a published classifier [11]

Top Ranked Proteins	The number of urinary biomarkers included[1]	P-value[1]	The number of urinary biomarkers included[2]	P-value[2]
500	12	0.0045	29	1.50e-11
1,000	21	0.0012	56	6.45e-21
1,500	34	1.15e-05	74	1.60e-24
2,000	48	2.60e-08	91	4.22e-28
2,500	68	2.83e-14	112	3.88e-35
3,000	86	4.24e-20	123	1.83e-35
3,500	106	3.74e-28	140	1.13e-40
4,000	113	6.83e-28	146	3.28e-38
4,500	116	3.01e-25	154	1.21e-37
5,000	120	1.34e-23	164	8.01e-39
5,500	126	2.06e-23	176	8.37e-42
6,000	129	1.43e-21	182	1.18e-40

[1] by using the classifier in a previous study [11]; [2] by using our classifier.

3.3 The Prediction of origins of urinary proteins

The prediction of urological origins of the predicted excretory proteins.

We have developed a classifier for predicting the urological origin of excretory proteins. The training of the classifier was done on a set of 430 proteins known to be of urological origin and 365 proteins known to be not of urological origin. An SVM-based classifier was trained along with a feature selection procedure based on the same set of 39 features totaling 1,537 dimensions (see Section 3.1), which gives rise to 111 final parameters.

Then 10-fold cross-validation was applied to the training set to evaluate the prediction performance of excretory proteins of urological origin. The performance by the trained classifier using a linear and RBF kernel, respectively, is shown in Table 4. Figure 2 shows the ROC curves.

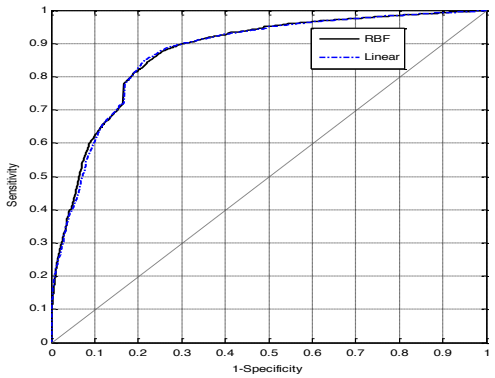

Fig. 2. The ROC curves for predicted urological origins of excretory proteins.

Table 4. The average performance of the prediction of urological origins for excretory proteins assessed by 10-fold cross validation

SVM Kernel	Sensitivity	Specificity	Precision	Accuracy	MCC	AUC
Linear	83.10%	77.90%	81.72%	80.72%	0.611	0.873
RBF	83.39%	77.90%	81.77%	80.88%	0.614	0.875

For the top 5,000 human proteins that have been predicted as excretory proteins in urine, we predicted that 2,357 are of urological origin and 2,643 are not. The function enrichment analysis of Gene Ontology and Pathway is used to understand the cellular functions and subcellular locations of the 2,357 urological origin proteins. The most significantly enriched biological processes, cellular components, and molecular function are biological adhesion, extracellular region, and GTP binding. Meanwhile, the most significantly enriched pathways are Glycolysis/Gluconeogenesis and Pyruvate metabolism (see Table S5). The function enrichment analysis of Gene Ontology and Pathway is used to understand the cellular functions and subcellular locations of the 2,643 non-urological origin proteins. The most significantly enriched biological pro-

cesses, cellular components, and molecular function are immune response, intrinsic to plasma membrane, and hormone activity. Meanwhile, the most significantly enriched pathways are Cytokine-cytokine receptor interaction and Complement and coagulation cascades (see Table S5). We can see that the significantly enriched pathways of urological origin proteins are related to metabolic pathways, and the significantly enriched pathways of non-urological origin proteins are related to immune systems and tissue repair.

3.4 Identification of urinary biomarkers for lung cancer.

We have applied the methods presented above to the gene-expression data of lung adenocarcinoma and squamous cell carcinoma, with an aim to predict urinary markers for the disease. By examining 102 lung cancer tissue versus 102 matching control tissues in the TCGA database [25], 5,491 genes are found to be differentially expressed in the cancer versus the control tissues. Using the prediction method given in Section 2.2, 587 of these genes are predicted to be urine excretory. Out of these proteins, 116 have been identified in human urines, including 13 that have been reported as potential urine biomarkers for non-small-cell lung carcinoma [29].

Table 5. Proteins predicted as urinary biomarkers of two types of lung cancer.

Not included in the training dataset			
Accession	**Protein Name**	**Ratio (cancer/normal)**	**D-value**
Q6ZMP0	Thrombospondin type-1 domain-containing protein 4	-1.74	1.466
P28908	Tumor necrosis factor receptor superfamily member 8	-1.92	1.431
P08833	Insulin-like growth factor-binding protein 1	1.45	1.591
O43240	Kallikrein-10	-15.05	1.110
P01127	Platelet-derived growth factor subunit B	-2.13	1.324
P39900	Macrophage metalloelastase	87.24	1.119
Included in the training dataset			
Accession	**Protein Name**	**Ratio (cancer/normal)**	**D-value**
P13688	Carcinoembryonic antigen-related cell adhesion molecule 1	2.23	2.401
P01033	Metalloproteinase inhibitor 1	1.62	1.479
P39060	Collagen alpha-1(XVIII) chain	1.21	1.922
P39059	Collagen alpha-1(XV) chain	2.34	1.238
P01024	Complement C3	-5.20	2.131
P10909	Clusterin	-4.19	1.791
P04085	Platelet-derived growth factor subunit A	-1.00	1.000

4 Discussions and conclusion

Early diagnosis plays a vital role in controlling diseases. Identifying disease-informing biomarkers represents an effective way for early diagnosis of a disease. The key is to identify the most useful biomarkers for disease detection. With the rapid development of omic technologies, a variety of disease tissue omic data are being generated and stored into publicly available databases. These data provided unprecedented opportunities to computational data analysts to develop effective methods to discover the most effective biomarkers for specific diseases.

Comparable to the existing biomarker prediction methods, our study has two novel aspects: (i) a new method for predicting if a cellular protein is urine excretory based on unique features of proteins known to be urine excretory; and (ii) a novel method for identifying urinary proteins originated from the urinary system. We anticipate that these ideas and methods will ultimately lead to substantially improved abilities for reliable identification of urinary biomarkers.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 81320108025, 61402194, 61572227), Development Project of Jilin Province of China (20140101180JC) and China Postdoctoral Science Foundation (2014T70291).

Reference

1. Lee, S., Huang, H. & Zelen, M. (2004) Early detection of disease and scheduling of screening examinations, *Stat Methods Med Res.* **13**, 443-56.
2. Thongboonkerd, V. (2007) Practical points in urinary proteomics, *J Proteome Res.* **6**, 3881-90.
3. Shao, C., Li, M., Li, X., Wei, L., Zhu, L., Yang, F., Jia, L., Mu, Y., Wang, J., Guo, Z., Zhang, D., Yin, J., Wang, Z., Sun, W., Zhang, Z. & Gao, Y. (2011) A tool for biomarker discovery in the urinary proteome: a manually curated human and animal urine protein biomarker database, *Mol Cell Proteomics.* **10**, M111 010975.
4. Abogunrin, F., O'Kane, H. F., Ruddock, M. W., Stevenson, M., Reid, C. N., O'Sullivan, J. M., Anderson, N. H., O'Rourke, D., Duggan, B., Lamont, J. V., Boyd, R. E., Hamilton, P., Nambirajan, T. & Williamson, K. E. (2012) The impact of biomarkers in multivariate algorithms for bladder cancer diagnosis in patients with hematuria, *Cancer.* **118**, 2641-50.
5. Raimondo, F., Morosi, L., Corbetta, S., Chinello, C., Brambilla, P., Della Mina, P., Villa, A., Albo, G., Battaglia, C., Bosari, S., Magni, F. & Pitto, M. (2013) Differential protein profiling of renal cell carcinoma urinary exosomes, *Mol Biosyst.* **9**, 1220-33.
6. Malavaud, B., Salama, G., Miedouge, M., Vincent, C., Rischmann, P., Sarramon, J. P. & Serre, G. (1998) Influence of digital rectal massage on urinary prostate-

specific antigen: interest for the detection of local recurrence after radical prostatectomy, *Prostate.* **34**, 23-8.
7. Ghoniem, G., Faruqui, N., Elmissiry, M., Mahdy, A., Abdelwahab, H., Oommen, M. & Abdel-Mageed, A. B. (2011) Differential profile analysis of urinary cytokines in patients with overactive bladder, *Int Urogynecol J.* **22**, 953-61.
8. Abdullah-Soheimi, S. S., Lim, B. K., Hashim, O. H. & Shuib, A. S. (2010) Patients with ovarian carcinoma excrete different altered levels of urine CD59, kininogen-1 and fragments of inter-alpha-trypsin inhibitor heavy chain H4 and albumin, *Proteome Sci.* **8**, 58.
9. Li, Y., Zhang, Y., Qiu, F. & Qiu, Z. (2011) Proteomic identification of exosomal LRG1: a potential urinary biomarker for detecting NSCLC, *Electrophoresis.* **32**, 1976-83.
10. Abdalla, M. A. & Haj-Ahmad, Y. (2012) Promising Urinary Protein Biomarkers for the Early Detection of Hepatocellular Carcinoma among High-Risk Hepatitis C Virus Egyptian Patients, *J Cancer.* **3**, 390-403.
11. Hong, C. S., Cui, J., Ni, Z., Su, Y., Puett, D., Li, F. & Xu, Y. (2011) A computational method for prediction of excretory proteins and application to identification of gastric cancer markers in urine, *PLoS One.* **6**, e16875.
12. Li, S. J., Peng, M., Li, H., Liu, B. S., Wang, C., Wu, J. R., Li, Y. X. & Zeng, R. (2009) Sys-BodyFluid: a systematical database for human body fluid proteome research, *Nucleic Acids Res.* **37**, D907-12.
13. Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C. H., Corthals, G. L., Costello, C. E., Deutsch, E. W., Domon, B., Hancock, W., He, F., Hochstrasser, D., Marko-Varga, G., Salekdeh, G. H., Sechi, S., Snyder, M., Srivastava, S., Uhlen, M., Wu, C. H., Yamamoto, T., Paik, Y. K. & Omenn, G. S. (2011) The human proteome project: current state and future direction, *Mol Cell Proteomics.* **10**, M111 009993.
14. Adachi, J., Kumar, C., Zhang, Y., Olsen, J. V. & Mann, M. (2006) The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins, *Genome Biol.* **7**, R80.
15. Li, Q. R., Fan, K. X., Li, R. X., Dai, J., Wu, C. C., Zhao, S. L., Wu, J. R., Shieh, C. H. & Zeng, R. (2010) A comprehensive and non-prefractionation on the protein level approach for the human urinary proteome: touching phosphorylation in urine, *Rapid Commun Mass Spectrom.* **24**, 823-32.
16. Marimuthu, A., O'Meally, R. N., Chaerkady, R., Subbannayya, Y., Nanjappa, V., Kumar, P., Kelkar, D. S., Pinto, S. M., Sharma, R., Renuse, S., Goel, R., Christopher, R., Delanghe, B., Cole, R. N., Harsha, H. C. & Pandey, A. (2011) A comprehensive map of the human urinary proteome, *J Proteome Res.* **10**, 2734-43.
17. Decramer, S., Gonzalez de Peredo, A., Breuil, B., Mischak, H., Monsarrat, B., Bascands, J. L. & Schanstra, J. P. (2008) Urine in clinical proteomics, *Mol Cell Proteomics.* **7**, 1850-62.
18. Hoorn, E. J., Pisitkun, T., Zietse, R., Gross, P., Frokiaer, J., Wang, N. S., Gonzales, P. A., Star, R. A. & Knepper, M. A. (2005) Prospects for urinary proteomics: exosomes as a source of urinary biomarkers, *Nephrology (Carlton).* **10**, 283-90.

19. Jia, L., Li, X., Shao, C., Wei, L., Li, M., Guo, Z., Liu, Z. & Gao, Y. (2013) Using an isolated rat kidney model to identify kidney origin proteins in urine, *PLoS One.* **8**, e66911.
20. Wood, S. L., Knowles, M. A., Thompson, D., Selby, P. J. & Banks, R. E. (2013) Proteomic studies of urinary biomarkers for prostate, bladder and kidney cancers, *Nat Rev Urol.* **10**, 206-18.
21. Cui, J., Liu, Q., Puett, D. & Xu, Y. (2008) Computational prediction of human proteins that can be secreted into the bloodstream, *Bioinformatics.* **24**, 2370-5.
22. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L., Tate, J. & Punta, M. (2014) Pfam: the protein families database, *Nucleic Acids Res.* **42**, D222-30.
23. Storey, J. D. & Tibshirani, R. (2003) Statistical significance for genomewide studies, *Proc Natl Acad Sci U S A.* **100**, 9440-5.
24. Scholkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. (2001) Estimating the support of a high-dimensional distribution, *Neural Computation.* **13**, 1443-1471.
25. Hampton, T. (2006) Cancer genome atlas., *JAMA.* **296**, 1958-1958.
26. Harris, N. S. & Winter, W. E. (2012) *Multiple myeloma and related serum protein disorders: an electrophoretic guide*, Demos Medical Publishing.
27. Consortium, U. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Research.* **40**, D71-D75.
28. Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. & Lempicki, R. A. (2003) DAVID: database for annotation, visualization, and integrated discovery, *Genome Biol.* **4**, P3.
29. Nolen, B. M., Lomakin, A., Marrangoni, A., Velikokhatnaya, L., Prosser, D. & Lokshin, A. E. (2014) Urinary protein biomarkers in the early detection of lung cancer, *Cancer Prev Res (Phila)*.