EBioMedicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Proof-of-Concept Workflow for Establishing Reference Intervals of Human Urine Proteome for Monitoring Physiological and Pathological Changes

Wenchuan Leng [a,1], Xiaotian Ni [a,d,1], Changqing Sun [b,1], Tianyuan Lu [a], Anna Malovannaya [c], Sung Yun Jung [c], Yin Huang [a,e], Yang Qiu [a], Guannan Sun [a], Matthew V. Holt [c], Chen Ding [a,b], Wei Sun [a], Xuebo Men [b], Tieliu Shi [d], Weimin Zhu [a,b], Yi Wang [c], Fuchu He [a], Bei Zhen [a,b,*], Guangshun Wang [b,**], Jun Qin [a,b,c,*]

[a] State Key Laboratory of Proteomics, National Center for Protein Sciences (The PHOENIX Center, Beijing), Beijing Proteome Research Center, Beijing 102206, China
[b] Joint Center for Translational Medicine, Tianjin Baodi Hospital, Tianjin 301800, China
[c] Alkek Center for Molecular Discovery, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA
[d] Center for Bioinformatics, East China Normal University, Shanghai 200241, China
[e] School of Life Science and Technology, Tongji University, Shanghai 200092, China

ARTICLE INFO

Article history:
Received 25 November 2016
Received in revised form 20 March 2017
Accepted 20 March 2017
Available online xxxx

Keywords:
Reference intervals
Urine proteome
Cancer
Biomarker
Mass spectrometry

ABSTRACT

Urine as a true non-invasive sampling source holds great potential for biomarker discovery. While approximately 2000 proteins can be detected by mass spectrometry in urine from healthy people, the amount of these proteins vary considerably. A systematic evaluation of a large number of samples is needed to determine the range of the variations. Current biomarker studies often measure limited number of urine samples in the discovery phase, which makes it difficult to determine whether proteins differentially expressed between control and disease groups represent actual difference, or are just physiological variations among the individuals, leads to failures in the validation phase with the increased sample numbers. Here, we report a streamlined workflow with capacity of measuring 8 urine proteomes per day at the coverage of >1500 proteins. With this workflow, we evaluated variations in 497 urine proteomes from 167 healthy donors, establishing reference intervals (RIs) that covered urine protein variations. We demonstrated that RIs could be used to monitor physiological changes by detecting transient outlier proteins. Furthermore, we provided a RIs-based algorithm for biomarker discovery and validation to screen for diseases such as cancer. This study provided a proof-of-principle workflow for the use of urine proteome for health monitoring and disease screening.

1. Introduction

Urine is a commonly used biological fluid for discovery of disease markers, diagnostics, and health status monitoring. Urine presents several distinct advantages over blood. For example, its sampling is truly non-invasive, therefore can be repeated frequently; the urine proteome is also simpler than the plasma proteome and more amenable to proteomic analysis (An and Gao, 2015; Shao et al., 2011b). Proteins in urine originate from glomerular filtration of plasma and secretion of urogenital system (Pisitkun et al., 2004, 2006; Sun et al., 2005; Wang et al., 2006) and changes in urinary protein composition can reflect physiological and pathological status of the human body (Decramer et al., 2008; Wu and Gao, 2015).

Much effort has been made to characterize protein composition of urine using mass spectrometry (MS) during the last decade (Adachi et al., 2006; Kentsis et al., 2009; Khristenko et al., 2016; Marimuthu et al., 2011; Nagaraj and Mann, 2011; Sun et al., 2009; Thongboonkerd et al., 2002). Databases, such as Max-Planck Unified Proteome database (http://mapuproteome.com/) (Zhang et al., 2007), the Human Kidney and Urine Proteome Project (http://www.hkupp.org/) (Yamamoto et al., 2008), the Human Urinary Proteome Database (http://mosaiques-diagnostics.de/diapatpcms/mosaiquescms/front_content.php?idcat=257) (Coon et al., 2008), Urinary Protein Biomarker (UPB) database (http://www.mybiosoftware.com/upb-20130710-urine-protein-biomarker-database.html) (Shao et al., 2011a), and Urine Proteomics.org (http://urineproteomics.org/databases.html) (Kentsis et al., 2009), documented lists of urinary proteins, providing convenient resources for keeping track of published urine proteomes. However, none of these databases provided quantitative information about the urine proteins.

* Corresponding authors at: State Key Laboratory of Proteomics, National Center for Protein Sciences (The PHOENIX Center, Beijing), Beijing Proteome Research Center, Beijing 102206, China.
** Correspondence to: G. Wang, Joint Center for Translational Medicine, Tianjin Baodi Hospital, Tianjin 301800, China.
E-mail addresses: zp1963@sina.com (B. Zhen), wgsTMUBH@163.com (G. Wang), jqin@bcm.edu (J. Qin).
[1] Equal contributing authors.

http://dx.doi.org/10.1016/j.ebiom.2017.03.028

A number of clinical proteomics studies have generated a long list of candidate urine protein biomarkers of various diseases (Beretov et al., 2015; Goodison et al., 2009; Rodriguez-Suarez et al., 2014; Shao, 2015); however, no biomarkers derived from 'discovery' studies were successfully translated into clinical practice to influence patient care and management (Fuzery et al., 2013; Mischak et al., 2010). One reason for these translational failures was the small sample size included in the studies, which did not have enough statistical power to distinguish between the difference resulting from pathological changes and the physiological variations among different individuals (Rifai et al., 2006; Rodriguez-Suarez et al., 2014).

At present, proteomics studies have revealed that intra-personal variation of urine proteomes is substantial (Nagaraj and Mann, 2011); the variation is further compounded by normal geno-proteomic differences among individuals. However, due to the lack of systematic evaluation of variations in human urine proteomes based on large population or long time span of sampling, it is still not clear whether these variations reflect fluctuation within a certain range in healthy persons and further, whether it is feasible to establish a protein reference range for the human urine proteome.

In clinical diagnosis, a normal range for a laboratory test is customarily established by a reference interval (RI) for its distributions in the population. Measurements of hundreds to thousands of people are required to make reliable empirical estimates (CLSI, 2010; Thompson et al., 2004). Typically, 2.5th and 97.5th percentiles of RI, which cover 95% of reference population are used as lower and upper limits, where outlier values are thought to signal potential problems for the test subject (CLSI, 2010). This approach is effective and widely used in diagnostics. But the concept has not been adopted in proteomics, as it requires the measurement of large number of samples.

Here we report a streamlined workflow to measure urine proteome from high-speed sediment of urine at the level of >1500 proteins within 3 h of MS running time. We measured 497 samples from 167 healthy donors, enabling us to evaluate day-to-day and inter-personal variations of the human urine proteome in a two-center setting. This dataset allowed us to establish intra-personal and pan-human RIs that covered variations of the human urine proteome. We presented examples using these RIs to identify outlier proteins that associated with physiological or pathological states, which might be used for health monitoring. Our study paved a way for biomarker discovery and validation for disease diagnosis and health monitoring by using urine proteome.

2. Materials and Methods

2.1. Sample Collection and Preparation

Midstream of the first-morning urine was obtained and stored at −80 C. Informed consents were signed by all test subjects and the study was approved by the Institutional Review Boards, Baylor College of Medicine (BCM) and Beijing Proteome Research Center (BPRC), respectively. Research adhered to the standards indicated by the Declaration of Helsinki. We used 10–20 ml of urine samples to establish the method. After establishing the standard operation procedure (SOP), 20 ml was used in the remaining experiments. Twenty milliliters of urine samples were centrifuged at 200,000 *g* for 70 min to save the pellets. We used a previously described method (Pisitkun et al., 2004) with modifications to remove uromodulin (*UMOD*; GeneID 7369). Briefly, 400 μl of resuspension buffer (50 mM Tris, 250 mM sucrose, pH 8.5) and dithiotheitol (DTT) was added to the pellets to a final concentration of 50 mM and the suspension was then heated at 65 °C for 30 min. Then wash buffer (10 mM TEA, 100 mM NaCl, pH 7.4) was added and a second ultracentrifugation was carried out for 30 min. The sediments were dissolved in sodium dodecyl sulfate (SDS) buffer (1 % SDS, 50 mM Tris, pH 8.5) and half of the samples were used for SDS-PAGE. Resolved proteins were visualized with Coomassie Brilliant Blue and 6 gel pieces were subjected to in-gel digestion by trypsin as previously described (Malovannaya et al., 2010). Sample metadata were summarized in Supplementary Table 1.

2.2. NanoHPLC-MS Analysis

The extracted peptides were re-suspended in 20 μl of loading solution (5 % methanol containing 0.1 % formic acid) and 5 μl was analyzed. Thermo Fisher Q Exactive and LTQ Orbitrap VelosPro coupled to nLC-1000 were used. A homemade trap column (2 cm × 75 μm) and an analytical column (10 cm × 75 μm), both packed with Reprosil-Pur Basic C18 (3 μm, Dr. Maisch GmbH, Germany) were used. A 75 min gradient of 5–28 % acetonitrile at a flow rate of 400 nl/min was used for on-line HPLC-MS.

For Q Exactive, the full MS scan range was set to 375–1300 *m/z* and trap size for MS1 and MS2 were 3×10^6 and 2×10^5, respectively. The mass resolution for MS1 and MS2 were 140,000 and 17,500 respectively. The top 25 ions were selected for higher energy collision dissociation (HCD) with collision energy set at 27. For LTQ Orbitrap VelosPro, the full MS scan range was set to 375–1300 *m/z* and trap size for MS1 and MS2 were 3×10^6 and 3×10^4, respectively. The mass resolution for MS1 was 100,000. The top 20 ions were selected for collision induced dissociation (CID) with collision energy set at 29. Dynamic exclusion was used after 1st identification with 10 s repeat duration and 30 s exclusion duration.

2.3. Protein Identification and Label-free Quantification

Proteome Discoverer (PD, V1.4, ThermoFisher) with Mascot (Mascot V2.3, Matrix Science) was used to search raw data against Human RefSeq database (the 2013.07.04). Mass tolerance for precursor ions was set to 20 ppm; mass tolerances of fragment ions were 0.02 and 0.5 Da for Q Exactive and LTQ Orbitrap VelosPro, respectively. Carbamidomethylation of cysteine, oxidation of methionine, acetylation of protein N-terminal were included as variable modifications. A maximum of one missed cleavages was allowed. All assigned peptides were filtered with 1 % false discovery rate (FDR) at peptide level. We only kept identifications with ≥2 unique peptides (1 % FDR and ion score > 20), which was stricter than 1 % FDR at the protein level.

All identified peptides were quantified with peak areas derived from their MS1 intensity. The process was as followed: 1) MS raw data were converted to the MS-platform independent mzXML format; 2) the spectral assignments from PD1.4 were then channeled through an in-housed pipeline to construct Extracted Ion Chromatogram (XIC) peaks with their corresponding intensity values included in mzXML data. For protein quantification, intensity based absolute quantification (iBAQ) algorithm (Schwanhausser et al., 2011) was used. To normalize the differences in loading amounts among samples, we then converted iBAQ value to FOT (fraction of total) - iBAQ value of each protein divided by the sum of all iBAQ values of all proteins in the sample. Thus, FOT number is a relative concentration for the protein in the total measurable proteome. FOTs of most proteins in a sample were very small and more than five decimal values were common. These small numbers would be visually difficult for human eyes. Therefore, we multiplied the FOT number with 10^5 to obtain iFOT5 to make easier visualization of values. All missing values were substituted with zero.

2.4. An algorithm for Screening Cancer

We first only kept proteins with ≥2 strict peptides (1 % FDR and ion score > 20). In the rest 450 normal samples, we randomly selected 350 samples and calculate their RIs based on the iFOT5 values. We used non-parametric 99.5th percentile values as the upper limits for selecting outlier proteins. We then randomly selected 45 cancer samples as the training data set to find outlier proteins that are outside of the RI upper limits, resulting in ~500 proteins. We then applied the same scheme on the validation dataset and to obtain the outlier pools, we then calculated the *p*-value for an overlap between the cancer outlier

Please cite this article as: Leng, W., et al., Proof-of-Concept Workflow for Establishing Reference Intervals of Human Urine Proteome for Monitoring Physiological and Pathologic..., EBioMedicine (2017), http://dx.doi.org/10.1016/j.ebiom.2017.03.028