Discovery and Multi-center Verification of Prostate Cancer Protein Biomarkers using Single-shot Short Gradient Microflow SWATH and MRM^{HR} Mass Spectrometry

Rui Sun 1,2*, Christie Hunter 3*#, Chen Chen 4, Weigang Ge 1,2, Nick Morrice 3, Qiushi Zhang 1,2, Xue Cai 1,2, Bo Wang 5, Xiaoyan Yu 6, Xiaodong Teng 5, Lirong Chen 6, Shaozheng Dai 7, Jian Song 8, Zhongzhi Luan 7, Changbin Yu 8, Ruedi Aebersold 9, Yi Zhu 1,2#, Tiannan Guo 1,2 #

1, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China

2, Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China

3, SCIEX, USA

4, SCIEX, China

5, Department of Pathology, The First Affiliated Hospital of College of Medicine, Zhejiang University, Hangzhou, China

6, Department of Pathology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, 310009, China

7, School of Computer Science and Engineering, Beihang University, Beijing, China

8, School of Engineering, Westlake University, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China

9, Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Switzerland

* co-first

co-correspondence

Emails:

Rui Sun: sunrui@westlake.edu.cn 0000-0002-8484-8089

Christie Hunter: Christie.Hunter@sciex.com 0000-0003-2587-1489

Chen Chen: chen.chen@sciex.com

Weigang Ge: geweigang@westlake.edu.cn 0000-0002-1803-4327

Nick Morrice: nick.morrice@sciex.com

Qiushi Zhang: zhangqiushi@westlake.edu.cn 0000-0001-5169-2092

Xue Cai: caixue@westlake.edu.cn 0000-0002-7427-2100

Bo Wang: 1506128@zju.edu.cn

Xiaoyan Yu: greg.sander15@yahoo.com

Xiaodong Teng: 1102069@zju.edu.cn

Lirong Chen: chenlr999@163.com

Shaozhi Dai: daiszemail@126.com 0000-0001-8901-1272

Jian Song: songjian@westlake.edu.cn

38 Zhongzhi Luan: luan.zhongzhi@buaa.edu.cn 0000-0002-7186-0556

39 Changbin Yu: yu_lab@westlake.edu.cn

40 Ruedi Aebersold: aebersold@imsb.biol.ethz.ch

41 Yi Zhu: zhuyi@westlake.edu.cn 0000-0003-0429-0802

42 Tiannan Guo: guotiannan@westlake.edu.cn 0000-0003-3869-7651

43

Abstract (limit 250 words)

BACKGROUND: Discovery and verification of protein biomarkers in clinical specimens using mass spectrometry are inherently challenging and resource-intensive.

METHODS: Formalin-fixed paraffin-embedded tissue-biopsy samples from a prostate cancer patient cohort (PCZA, n = 68) were processed in triplicate using pressure cycling technology, followed by microflow LC SWATH® analysis with different gradients and window schemes. Potential protein biomarker candidates were prioritized using random forest analysis and evaluated by receiver operating characteristic curve analysis. Selected proteins were verified with a targeted MRM^{HR} assay using the 15 min microflow LC strategy on a second prostate cancer cohort (PCZB, n = 54). Potential biomarkers were further verified using TMA on a third cohort (PCZD, n = 100).

RESULTS: We developed and optimized a 15-min microflow LC approach coupled with microflow SWATH MS. Application of the optimal 15 min and conventional 120 min LC gradient scheme using samples from the PCZA cohort led to quantification of 3,800 proteins in both methods with high quantitative correlation (r = 0.77). MRM^{HR} verification of 154 prioritized proteins showed high quantitative consistency with the 15 min SWATH data (r = 0.89). Separation of benign and malignant tissues achieved precision (AUC = 0.99). ECHS1 was further verified in a third cohort PCZD successfully, agreeing with RNAseq data from the TCGA in a different cohort (n=549). Our methods enables practical proteomic analysis of 204 tissue samples within 5 working days.

CONCLUSION: Single-shot, short gradient SWATH-MS coupled with MRM^{HR} is both practical and effective in discovering and verifying protein biomarkers in clinical specimens.

Introduction

Prostate cancer (PCa) is a group of complex diseases that occur in the prostate gland, resulting from carcinogenesis that leads to the modulation of a number of proteins. The regulated proteins in malignant prostate tissue samples reflect the molecular pathology of these complex diseases and offer a plethora of drug targets[1]. Due to the high degree of inter-patient and inter-tissue heterogeneity, it is crucial to identify differentially expressed proteins in prostate cancer cohorts and verify them in multi-center studies. However, this is inherently challenging and resource-intensive mainly due to the lack of high-throughput protein biomarker discovery and verification methods.

Typically, proteomes are analyzed using shotgun proteomics by multi-dimensional fractionation based on the need to isolate each peptide precursor for fragmentation, sequencing, and quantification. Isotopic labeling technologies including SILAC, iTRAQ, and TMT increase the multiplexing capability of shotgun proteomics; however, they remain resource-intensive for analyzing clinical cohorts. Several proteomics studies using these methods have reportedly analyzed up to 200 clinical samples [2]; however, very few laboratories—even well-equipped ones—can carry out such expensive investigations. Many researchers have been striving to shorten the LC gradient time to increase the efficiency of shotgun proteomics analysis. For example, by integrating multiple pre-fractionations with relatively shorter LC gradients, in-depth MS analysis of a whole proteome can be achieved [3]. However, this analytical approach lacks reproducibility and is time consuming, posing a challenge in applying it to quantitative clinical studies for biomarker discovery for which hundreds to thousands of patients are usually recruited. A faster method based on Nano-LC and Orbitrap MS has been reported recently [4] that achieved high throughput, with 60 samples per working day with a 21-min nano-LC system. However, the robustness and reproducibility of this method for large numbers of tissue samples in cohort studies remain to be evaluated.

SWATH/DIA mass spectrometry [5], which identifies and quantifies peptide precursors via chromatographic peak groups from highly convoluted mass spectra [6], bypasses the need to isolate peptide precursors during acquisition, improving data completeness and enabling efficient single-shot proteomic analysis. The key to this MS technique is the ability to collect high resolution MS/MS spectra at very high acquisition rates, such that a wide mass range can be covered with a series of

smaller Q1 isolation windows in an LC-compatible cycle time. Thus, the rapid scanning rate of TripleTOF® systems has been the key to enabling the shortening of LC gradients for analyzing complex tissue proteomes, from 120 min [6] to 45 min [7], without compromising proteome depth; this method is increasingly applied to analyze various types of clinical samples including plasma [8] as well as tumor tissues [6, 7, 9].

An increasing number of studies have demonstrated the applicability of microflow chromatography coupled with SWATH-MS [10-14], especially with regard to clinical applications. In this study, we combined microflow LC with SWATH-MS to perform a quantitative proteomics study of PCa tissue-biopsies in a high throughput manner by further shortening the LC gradient—a method referred to as single-shot short gradient microflow SWATH acquisition (S3 SWATH MS). After benchmarking, we applied S3 SWATH to discover protein biomarker candidates from a prostate cancer cohort, some of which were further validated orthogonally using independent cohorts and MRM^{HR} verification.

Materials and Methods

Standard protein digests

HEK 293 cell digests were prepared as has been previously described [15] and were provided by Dr Yansheng Liu. K562 cell digests were obtained from the SWATH Performance Kit (SCIEX, Framingham, MA, USA). iRT peptides (Biognosys, Schlieren, Switzerland) were spiked into peptide samples at a final concentration of 10% prior to MS analysis for retention time (RT) calibration.

PCa patient cohorts and formalin-fixed paraffin-embedded (FFPE) samples

The PCZA and PCZB cohorts were acquired from the Second Affiliated Hospital College of Medicine, Zhejiang University. The PCZD cohort was collected from the First Affiliated Hospital College of Medicine, Zhejiang University. All patients were recruited from 2017 to 2018. All cohorts were approved by the ethics committee of their respective hospitals.

The PCZA cohort was composed of 58 PCa patients and 10 benign prostatic hyperplasia (BPH) patients. The PCZB cohort included 24 PCa patients and 30 BPH patients, while the PCZD cohort contained 70 PCa patients and 30 BPH patients.

In the PCZA cohort, three biological replicates (size $1 \times 1 \times 5$ mm^3) were collected and analyzed by SWATH MS and MRM^{HR}. In the PCZB cohort, two biological replicates ($1.5 \times 1.5 \times 5$ mm^3) were analyzed by MRM^{HR}. In the PCZD cohort, one punch ($1.5 \times 1.5 \times 5$ mm^3) was used for each sample for TMA validation.

Pressure cycling technology (PCT)-assisted peptide extraction from FFPE tissues

Approximately 0.5 mg of an FFPE tissue punch was weighed and processed for each biological replicate via the FFPE-PCT workflow as described previously [16]. Briefly, the tissue punches were first dewaxed by incubating with 1 mL of heptane under gentle vortexing at 600–800 rpm, followed by serial rehydration using 1 mL of 100%, 90%, and 75% ethanol, respectively. Tissues were further incubated with 200 µL of 0.1% formic acid (FA) at 30 °C for 30 min for acidic hydrolysis. The tissue punches were then transferred into PCT-MicroTubes and were briefly washed with 100 µL of fresh 0.1 M Tris-HCl (pH 10.0) to remove FA residues. Thereafter, the tissues were incubated with 15 µL of freshly prepared 0.1 M Tris-HCl (pH 10.0) at 95 °C for 30 min with gentle vortexing at 600 rpm. Samples were immediately cooled to 4 °C after basic hydrolysis.

Following the pretreatment described above, 25 µL of lysis buffer (6 M urea, 2 M thiourea, 5 mM Na_2EDTA in 100 mM ammonium bicarbonate, pH 8.5) was added to the PCT-MicroTubes containing tissues and protein extracts that were soaked previously in 15 µL of 0.1 M Tris-HCl (pH 10.0). The tissue samples were further subjected to PCT-assisted tissue lysis and protein digestion procedures using the Barocycler NEP2320-45K (Pressure Biosciences Inc., Boston, MA, USA) as

described previously [17]. The PCT scheme for tissue lysis was set such that each cycle involved 30 s of high pressure at 45 kpsi and 10 s of ambient pressure, oscillating for 90 cycles at 30 °C. Protein reduction and alkylation was performed at ambient pressure by incubating protein extracts with 10 mM Tris(2-carboxyethyl) phosphine (TCEP) and 20 mM iodoacetamide (IAA) in darkness at 25 °C for 30 min, with gentle vortexing at 600 rpm in a thermomixer. Then the proteins were digested with MS grade Lys-C (enzyme-to-substrate ratio, 1:40) using a PCT scheme set to 50 s of high pressure at 20 kpsi and 10 s of ambient pressure for each cycle, oscillating for 45 cycles at 30 °C. Thereafter, the proteins were further digested with MS grade trypsin (enzyme-to-substrate ratio, 1:50) using a PCT scheme with 50 s of high pressure at 20 kpsi and 10 s of ambient pressure in one cycle, oscillating for 90 cycles at 30 °C. Peptide digests were then acidified with 1% trifluoroacetic (TFA) to pH 2–3 and subjected to C18 desalting. iRT peptides were spiked into peptide samples at a final concentration of 10% prior to MS analysis for RT calibration.

Optimization of microflow LC gradients coupled with SWATH MS

During the optimization studies, peptide samples were separated with different microflow gradients and different SWATH-MS parameters. Linear gradients of 3–35% acetonitrile (0.1% formic acid) with durations of 5, 10, 20, 30, and 45 min were evaluated. The number of Q1 variable windows (40, 60, 100) and MS/MS accumulation time (15, 25 ms) constituted the key parameters that were adjusted for the shorter gradients. The need for collision energy spread with the optimized collision energy ramps was tested. Four replicates were performed for each test, after which the data were processed with PeakView® software with the SWATH 2.0 MicroApp to evaluate the number of proteins and peptides quantified at <1 % FDR and with < 20% CV. The optimized methods were then tested on multiple instruments with different cell lysates to confirm the robustness of the observations.

S3 SWATH MS acquisition

Peptides were separated at a flow rate of 5 µL/min along a 15 min 5–35% linear LC gradient (buffer A: 2% ACN, 0.1% formic acid; buffer B: 80% ACN, 0.1% formic acid) using an Eksigent NanoLC 400 System coupled to a TripleTOF® 6600 system (SCIEX). The DuoSpray Source was replumbed using 25 µm ID hybrid electrodes to minimize post-column dead volume. The SWATH method consisted of a 150 ms TOF MS scan with m/z ranging from 350 to 1250 Da, followed by MS/MS scans performed on all precursors (from 100 to 1500 Da) in a cyclic manner. A 100 variable Q1 isolation window scheme was used in this study (Supplemental Table 1B). The accumulation time was set at 25 ms per isolation window, resulting in a total cycle time of 2.7 s.

For the beta-galactosidase digest (β-gal) calibration analysis, peptides were separated at a flowrate of 5 µL/min along a 5 min 5–35% linear LC gradient (buffer A: 2% ACN, 0.1% formic acid; buffer B: 80% ACN, 0.1% formic acid) using an Eksigent NanoLC 400 System coupled to a TripleTOF® 6600 system (SCIEX). The DuoSpray Source was replumbed using 25 µm ID hybrid electrodes to minimize post-column dead volume. The MS method consisted of a 250 ms TOF MS scan with m/z ranging from 400 to 1250 Da, followed by a 500 ms product ion scan (target m/z = 729.4, indicating a peptide in the β-gal digest mixture) with a scan range of 100–1500, high sensitivity mode. The RT, intensity, and m/z of targeted precursor and fragment ions were used for LC QC, the sensitivity test, and mass calibration separately.

MRM^{HR} MS acquisition

A time scheduled MRM^{HR} targeted quantification strategy was used to further validate proteins observed to be differentially expressed based on the SWATH quantification described above. Peptides were separated at 5 µL/min using the same microflow LC approach as that used for S3 SWATH MS analysis. The TripleTOF® 6600 mass spectrometer was operated in IDA mode for time-scheduling the MS/MS acquisition for 286 peptides for the MRM^{HR} workflow. The method consisted of one 75 ms TOF-MS scan for precursor ions with m/z ranging from 350 to 1250 Da, followed by MS/MS scans for fragment ions with m/z ranging from 100 to 1500 Da, allowing for a maximum of 45 candidate ions to monitor per cycle (25 ms accumulation time, 50 ppm mass tolerance, rolling collision energy, +2 to +5 charge states with intensity criteria above 2 000 000 cps to guarantee that no untargeted

peptides would be acquired). The fragment information including m/z and RT of a targeted precursor ion was confirmed by previous SWATH results and was then added to the inclusion list for the targeted analysis. The intensity threshold of targeted precursors in the inclusion list was set to 0 cps and the scheduling window was 60 s. The targeted peptide sequences were the same as those found in the previous SWATH MS analysis.

Targeted MRM^{HR} data were analyzed by Skyline [18], which automatically detected the extracted-ion chromatogram (XIC) from an LC run by matching the MS spectra of the targeted ion against its spectral library generated from the IDA mode within a specific mass tolerance window around its m/z. All peaks selected were checked manually after automated peak detection using Skyline. Both MS1 and MS2 filtering were set as "TOF mass analyzer" with a resolution power of 30 000 and 15 000, respectively, while the "Targeted" acquisition method was defined in the MS/MS filtering.

SWATH data analysis

The optimization data for optimal LC gradients were processed using the SWATH 2.0 MicroApp in PeakView® software (SCIEX) using the Pan Human Library [19]. RT calibration was performed first using iRT peptides for an adjusted RT window at a 75 ppm XIC extraction width. Replicate analysis was performed using the SWATH Replicate Analysis Template (SCIEX) to determine the number of peptides and proteins quantified at a 1% peptide FDR and < 10 or 20% CV.

The prostate samples were processed using the OpenSWATH pipeline. Briefly, SWATH raw data files were converted in profile mode to mzXML using msconvert and analyzed using OpenSWATH (2.0.0) [5] as described previously [6]. The RT extraction window was 600 s and m/z extraction was performed with 0.03 Da tolerance. RT was then calibrated using both iRT peptides. Peptide precursors were identified by OpenSWATH and PyProphet with d_score > 0.01. For each protein, the median MS2 intensity value of peptide precursor fragments, which were detected to belong to the protein was used to represent the protein abundance.

Tissue microarray analysis

The TMA and IHC procedures used in this study have been described previously [20]. The ESCH antibody was acquired from Proteintech (66489-1-Ig; Chicago, IL, USA).

Results and Discussion

Establishment of the S3 SWATH MS method

We first optimized the short microflow LC gradient using the TripleTOF® 6600 system. A series of LC gradient lengths ranging from 5, 10, 20, and 45 min were tested (**Table S1**). For each LC condition, the number of variable Q1 windows (60 and 100) were investigated. More acquisition windows require shorter MS/MS accumulation times to maintain an LC compatible cycle time; therefore, two different MS/MS accumulation times, 15 and 25 ms, were tested (**Fig. 1a**). These adjustments were aimed to ensure reasonable data points across the narrower chromatography peaks that are obtained with faster LC gradients.

Standard HEK 293 cell lysate digests were used for the optimization. One microgram of peptides was loaded onto the microflow column. For each optimization condition, four technical replicates were analyzed. The data were processed and evaluated according to the number of proteins and peptides quantified with FDR < 1% and CV < 10% or CV < 20%, respectively. The data acquired with 20 min microflow LC gradient enabled the quantification of 90% of the proteins as quantified by the 45 min microflow LC gradient method (**Fig. 1a**). Using the optimized 20 min gradient method, we identified 14,112 peptides and 3523 proteins with CV below 10% when the Q1 window was fixed at 100. Decreasing the MS/MS accumulation time to 15 ms was tested and did contribute to higher numbers of protein and peptide identifications, especially for the 5 and 10 min LC gradients. Calculation of the number of peptides identified per time unit (min) further showed that the LC-MS

condition involving a 10 min microflow LC gradient plus 100 Q1 windows with the MS/MS accumulation time of 15 ms resulted in the highest number of identified peptides per time unit. Typically, using more variable Q1 isolation windows provided improved quantitative results most likely because of reduced signal/noise due to increased coelution of peptides. The MS/MS accumulation time was shortened to account for the much narrower LC peak widths.

After the optimization experiments, the best condition for each gradient length with optimal windows and MS/MS accumulation time was used to analyze the HEK293 cell line proteome on multiple instruments to fully characterize the impact of gradient length on SWATH quantitation (**Fig. 1b**). As expected, as the gradients were shortened, fewer proteins and peptides could be reliably quantified. The 20 min gradient using the best acquisition conditions quantified ~91% and ~83% of the proteins and peptides, respectively, that were quantified during the 45 min gradient with a 1 µg peptide load. The 10 min gradient using the best acquisition conditions quantified ~77% and ~65 % of the proteins and peptides, respectively, that were quantified during the 45 min gradient with a 1 µg peptide load. A total of 3354 proteins were reproducibly quantified at < 20% CV using the 10 min gradient (**Fig. 1b**). That said, the reduction in quantified proteins was less than expected when compared to the 45 min gradient with a 1 µg peptide load. These results confirmed that accelerated microflow SWATH experiments, i.e., the S3 SWATH method, can be applied to biomarker research studies where large cohorts are available and high throughput is desired.

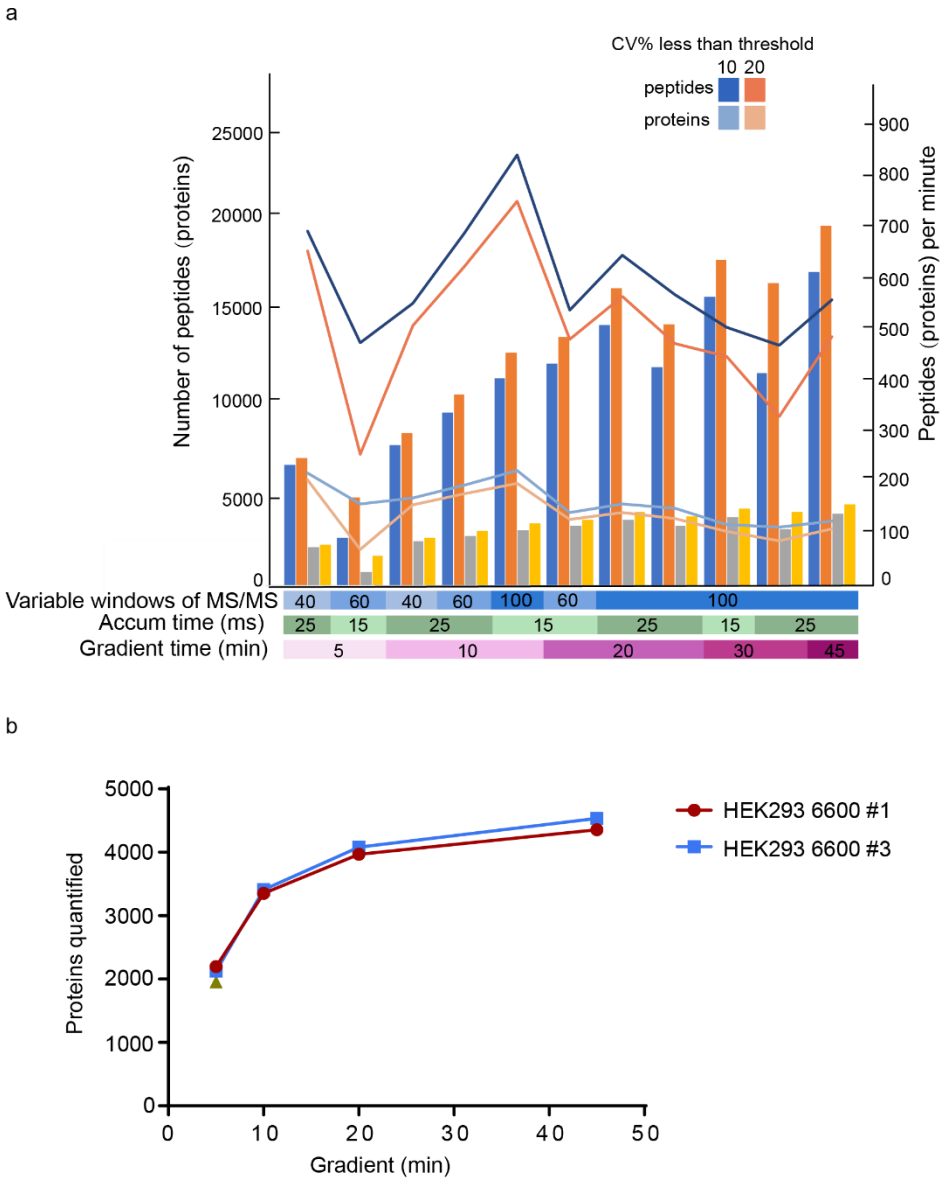

Figure 1. Optimization of S3 SWATH acquisition using HEK 293 cell lysate. **(a)** Using a series of microflow LC gradient lengths, the SWATH acquisition conditions were optimized to determine the best acquisition and processing settings for the accelerated analysis. The number of proteins and peptides were quantified with the cutoff of FDR < 1% and CV < 10% or < 20% calculated from four technical replicates for each tested condition. **(b)** Using the optimized conditions determined for each gradient, several datasets were collected on two 6600 instruments to benchmark the impact on proteins quantified. The 10 and 20 min gradients allowed for ~70% and ~90%, respectively, of the proteins to be quantified as compared to the 45 min gradient.

To confirm the utility of S3 SWATH in a PCa patient cohort, we further tested the S3 SWATH method using a PCa tissue pool which included three PCa and three BPH patient samples. We investigated the best MS acquisition and processing settings with each LC gradient of 10, 15, and 20 min gradients using a 1 µg peptide load in technical duplicate (**Fig. S1**). The 45 min microflow LC gradient with a 1 µg peptide load was used as reference. The 15 min gradient was selected for use in the study, as there were 3263 proteotypic peptides and 1367 SwissProt protein groups quantified with CV% below 10% for the 1 µg peptide load. As the use of shorter gradients is expected to have an effect on peak widths, S/N of data, etc., we also varied the RT extraction windows in the OpenSWATH analysis from 0 to 700 with a 10 s

interval and checked the protein identification (**Fig. S2**). Data showed that for both 10 min and 20 min gradient S3-SWATH, the number of proteins identified saturated when the RT extraction window was higher than 100 s.

Application of S3 SWATH to a PCa patient cohort

With the optimal S3 SWATH workflow, we analyzed the proteomes of the PCZA cohort containing 58 PCa patient samples and 10 BPH patient samples. The demographic and clinical information of these 68 patients are provided in **Table 1**; more details are available in **Table S2**. Altogether, we processed 204 FFPE tissue punches (three biological replicates for each sample) in seven batches (**Fig. 2a**), as well as quality control (QC) samples in each batch for PCT-assisted digestion (**Fig. 2a**). These samples were analyzed with S3 SWATH on a TripleTOF® 6600 mass spectrometer with QC runs for each MS batch. Automated β-gal calibration and analytical column washing were performed every four sample injections throughout the process (**Fig. 2b**). Taking into account the β-gal calibration, column washing time, and control samples, the S3 SWATH method with 100 variable Q1 windows on the TripleTOF® 6600 system completed the data acquisition for this cohort—in triplicates—in 125.7 hr (~5 days). We compared the S3 SWATH application in PCZA with another study by our research group that involved use of the conventional SWATH method with 120 min LC gradient and 48 variable Q1 windows in a TripleTOF® 5600+ [16]. The conventional method required 467 hr (~20 days) to analyze the sample cohort (**Fig. 2b**). A total of 5059 and 4038 SwissProt proteins were quantified by the 120 and 15 min workflows, respectively, with a quantitative reproducibility of CV < 20%. A total of 3800 proteins were found in common in both data sets (**Fig. 2b**), comprising 75.1% of the proteins in the 120 min SWATH and 94.1% proteins in the S3-SWATH. The S3 SWATH method gained a practical acceleration of 3.7 times when compared to the conventional method, with only a 24.88% loss of protein identification.

Both methods achieved a high degree of quantitative reproducibility at the protein level (**Fig. 2c**); the correlation coefficient between triplicates in the S3 SWATH method alone and the 120 min SWATH MS method is 0.874 and 0.865, respectively, indicating the impressive robustness of the S3 SWATH method. As for the 3800 proteins quantified, Pearson correlation showed a high similarity ($r = 0.7681$) between the proteome data generated by the two methods (**Fig. 2d**) and the correlation coefficient between two MS methods for individual samples is over 0.7 (**Fig. S3**), further consolidating the high quantitative accuracy of the S3 SWATH method.

312 **Table 1. Demographic and clinical characteristics of the patients from different cohorts**

313

314

Cohorts	PCZA		PCZB		PCZD	
Tissue types	BPH=10	PCa=58	BPH=30	PCa=24	BPH=30	PCa=70
Biological replicates	n=3		n=2		n=1	
Age (yr)						
Median	68	67	69	70	70	67
Range	58-80	52-82	54-82	58-78	61-83	53-81
Gleason score						
Median		7		7		7
Range		6-10		7-9		6-9
3+3		0		0		0
3+4		15		11		27
4+3		19		8		30
⩾8		17		3		17

315

316

317

318

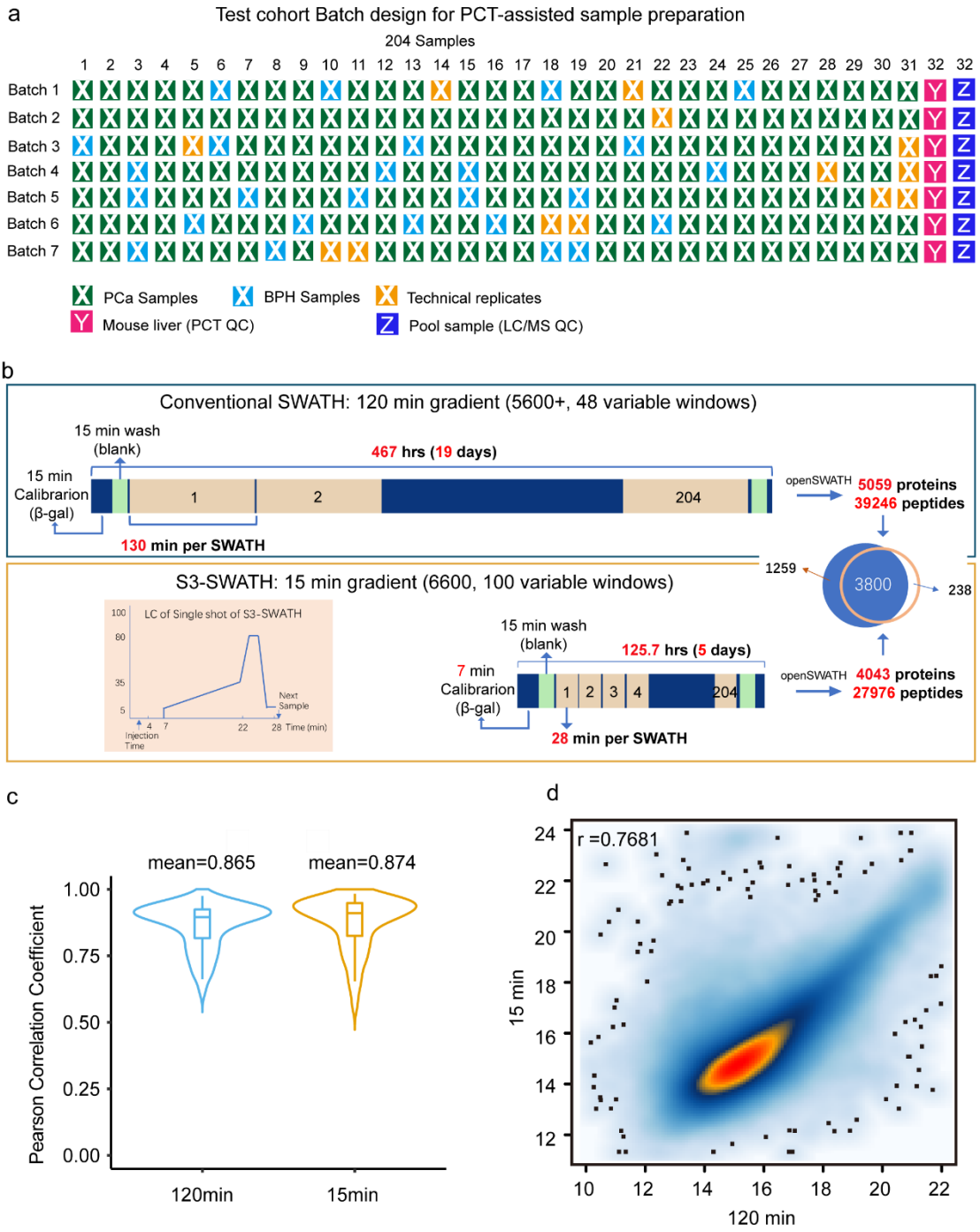

Figure 2. Application of S3 SWATH to the PCZA cohort. (a) Batch design of PCZA cohort. There are six batches of samples. Each batch contains PCa samples, BPH samples, technical replicates for randomly selected samples, mouse live-sample pool for PCT QC, and pooled prostate samples for LC-MS QC. **(b)** Workflow of the S3 SWATH and conventional SWATH. **(c)** Pearson correlation of 15 min and 120 min SWATH proteomes of the PCZA cohort. **(d)** Pearson correlation of 3800 proteins that were quantified by both 15 min and 120 min SWATH.

Next, we investigated the differentially expressed proteins using random forest (RF) analysis (**Fig. 3**). We selected 473 proteins from the S3 SWATH data set after filtering with Mean Decrease Accuracy > 0. We assessed the 120 min dataset further and found 152 proteins regulated in both data sets, which separated the malignant from the benign prostate samples in individual datasets with the area under the curve (AUC) value over 0.95.

Another classification algorithm, XGBoost, combined with the leave-one-out cross

validation, were used to validate the predictability of the 152-protein-signature panel. As shown in **Fig. 3**, we use the 120 min data and 15 min data as training set and validation set respectively. This result further demonstrated the robustness of the S3 SWATH workflow. Throughout the training process, due to the limited number of samples (n = 68), we used the leave-one-out cross validation method to judge the effect of the classifier on the training set and adjust the classifier parameters. The trained model then classified the test data set. Using the 120 min data set as the training set and the 15 min data set as the validation set, or vice versa, we observed high AUC using the XGBoost classifier (0.997 and 0.998 respectively).

Among the 152 proteins, many well-known diagnostic makers such as AMARCA or drug targets such as IDH, EIF4E [21] have been found. Next we employed ingenuity pathway analysis (IPA) to analyze the pathways represented by the 152 proteins (**Fig. S4-5**) and observed the presence of multiple metabolic pathways such as mitochondrial dysfunction, oxidative phosphorylation, fatty acid oxidation, and the TCA cycle, which is consistent with prior knowledge that metabolism is dysregulated and reprogrammed in prostate cancers [22] (**Fig. S4**). The IPA identified multiple drivers of the dysfunctional protein network observed here, including MYC, TP53 [23], and FOS [24], as well as potential drugs for PCa including decitabine [25], fenofibrate [26], and methotrexate [27] (**Fig. S5**).

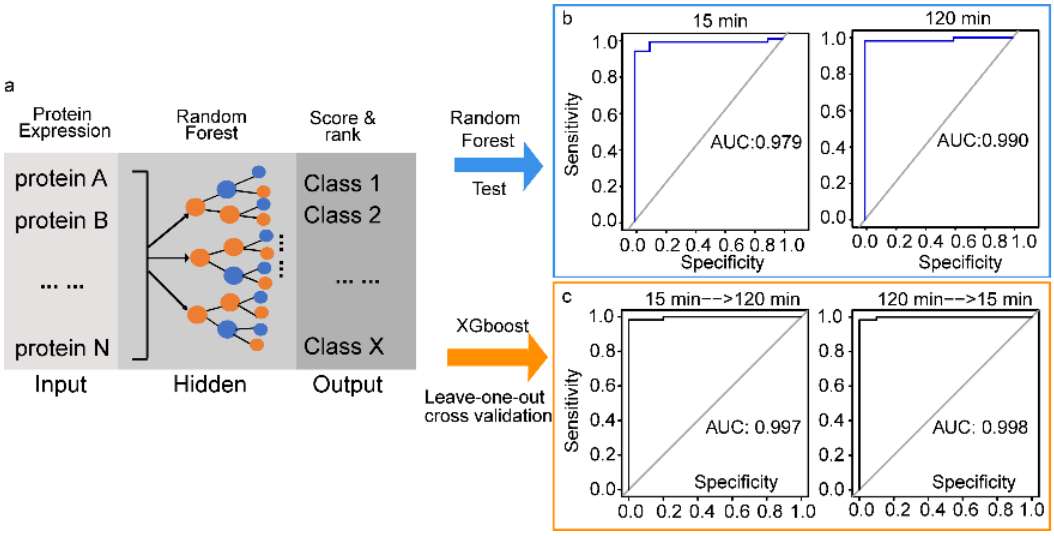

Figure 3. Identification of differentially expressed proteins between PCa and BPH tissues. (a) Schematic of the random forest algorithm. The protein matrix generation from 120 min SWATH and S3 SWATH involve independent data input and the output results represent the important score (mean decrease accuracy) for each protein. **(b)** A total of 152 overlapping proteins with higher scores were selected to distinguish benign and tumor tissues by AUC in two protein matrices from 120 min SWATH and S3-SWATH. **(c)** A total of 152 proteins were validated by another machinal learning method, XGBoost with leave-one-out cross validation, in two protein matrices.

Verification of potential diagnostic proteins using MRM^{HR}

We then further validated the S3 SWATH data using a PRM [28] implementation in the TripleTOF® system called MRM^{HR}. The MRM^{HR} method was optimized using a pooled prostate sample to select the best peptides, Q1 isolation windows, and best target fragment ions for quantitation. The protein and peptide information including the RTs were imported into Skyline to build a spectral library. Twelve of the original 152 proteins were rejected due to lower data quality. A total of 261 peptides from 140 proteins and 1429 fragment ions were selected for data extraction. Time scheduling was used to ensure at least eight data points across the LC peaks as well as an optimized accumulation time of 25 ms for each peptide for high quantitative data quality. We also examined the reproducibility of XICs for all peptides in the MRM^{HR} assays. For the five pooled samples measured across five batches, we found

that 76.6% of precursors measured from the peptides were quantified with a CV below 20%. The median CV was 13.4% (**Fig. S7a**).

To confirm the quantitative accuracy of the S3 SWATH data, we re-analyzed 99 samples in the PCZA cohort using the MRM^{HR} method. The protein fold-changes between tumor and normal samples were calculated. We investigated the correlation of 15 min MRM^{HR} and 15 min S3 SWATH quantitative datasets for the 140 proteins based on both Pearson and Spearman correlations. The two datasets were highly correlated with each other, confirming the superior accuracy of both the S3 SWATH and MRM^{HR} approaches (**Fig. 4a**). We further quantified the expression levels of the 140 proteins in an independent prostate cancer cohort, PCZB, containing 30 BPH and 24 PCa in duplicated biological replicates using the same 15 min MRM^{HR} workflow. For the six pooled samples measured across six batches, 75.6% of peptide precursors were quantified with a CV below 20%. The median CV is 14.9% (**Fig. S7b**). As shown in **Fig. 4b,** the ROC of the 140-protein-signature panel clearly distinguished PCa from BPH patient groups.

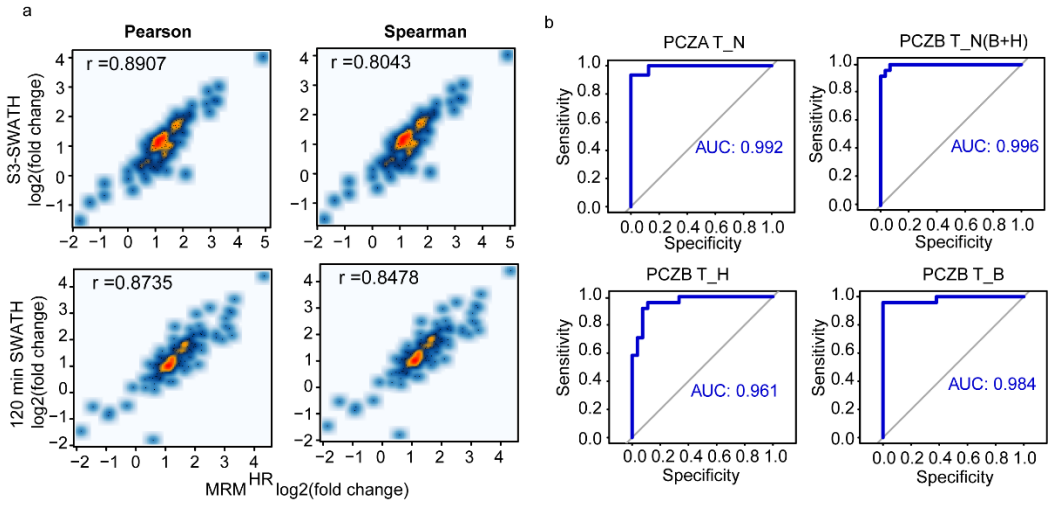

Figure 4. Verification of proteomic data using MRM^{HR}. (a) Pearson and Spearman correlation coefficients between 15 min S3 SWATH and MRM^{HR} datasets were calculated based on the $\log_2$(T/N) of protein expression. **(b)** The ROC curves of protein quantification from MRM^{HR} to predict the tumor and normal tissues with the random forest algorithm (T: PCa, N: BPH, H: hyperplasia in BPH patients, B: benign in BPH patients).

We then investigated three proteins—PRDX3 (P30048), COPA (P53621), and ECHS1 (P30084)—in detail, prioritized due to their participation in oncogene regulation and potential as drug targets (**Fig. S5**). PRDX3 is an androgen-regulated cell-surface protein which has been reported to be upregulated in PCa as a potential drug target [29, 30]. COPA is a coatomer mediating the biosynthetic protein transport from the endoplasmic reticulum and is associated with cell proliferation [31, 32]. It is overexpressed in PCa tissue and its inhibitor can suppress cell cycle and increase apoptosis of PCa [33]. ECHS1 is an enoyl-CoA hydratase in the mitochondria, which plays important roles in the mitochondrial fatty acid β-oxidation pathway with several reports associating it to HCC and cancers other than PCa [34-36]. Our data show that their expression changes appeared consistent in all three workflows, i.e., 120 min SWATH, S3-SWATH, and MRM^{HR} in the PCZA cohort samples (**Fig. 5**). An independent cohort, PCZB, further confirmed their upregulation in prostate tumors (**Fig. 5**). The ROCs of these three proteins from four different datasets distinguishing benign from malignant tissue samples are shown in **Fig. S9**; all of AUC were over than 0.75. The results from the independent cohort PCZB were better than those from cohort PCZA in terms of predictive power, probably due to the higher number of normal samples in the PCZB.

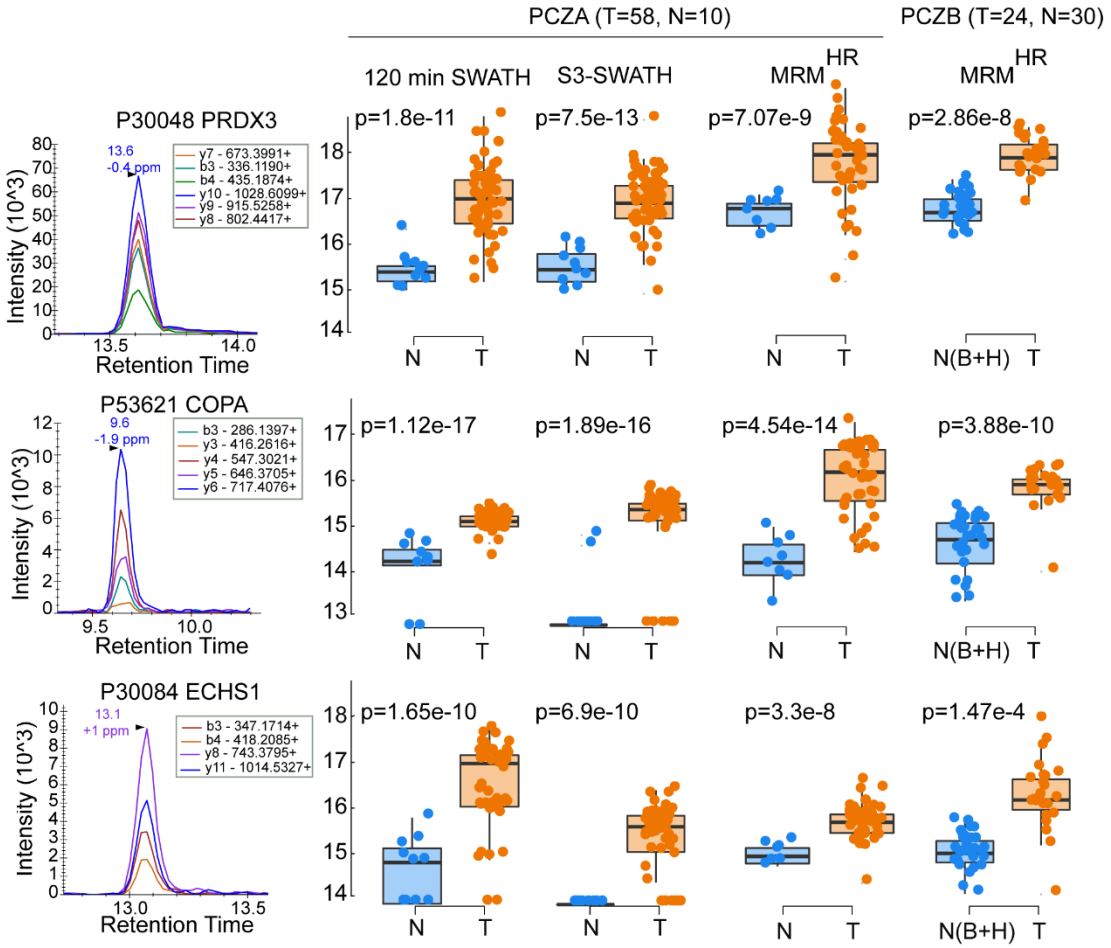

Figure 5. MRM^{HR} validation of potential diagnostic proteins using the PCZA and PCZB. PRDX3 (peptide: +2 DYGVLLEGSGLALR), COPA (peptide: +2 DVAVMQLR), ECHS1 (peptide: +2 EGMTAFVEK). The left panel shows the fragment ion extracted-ion chromatograms (XICs) for the peptide from each protein. The right panel of boxplots shows the peptides quantified in the different data sets.

Verification of protein markers using tissue microarray (TMA)

The most important disordered functions in prostate cancer involve metabolism. We performed IPA network analysis for the 140 proteins and identified 36 core networks (**Table S4**). Among them, Network No. 24, containing ECHS1, is associated with amino acid metabolism, small molecular biochemistry, and development (**Fig. 6a**). This network contains NF-KB signaling and the molecular associations of lipid metabolism [37], such as the participation of ACTA2 in the cholesterol biosynthesis pathway, which is involved in castration-resistant prostate cancer [38, 39]. ACTA2, ARSA, and ECHS1 are involved in fatty acid β oxidation. ECHS1 was found to be overexpressed in PCa tumors from both PCZA and PCZB cohorts, in all the four MS datasets in this study. We further applied TMA and immunohistochemistry staining (IHC) to verify the protein expression changes of ECHS1 measured by MRM^{HR} in the PCZD cohort which contains 30 BPH and 70 PCa (**Table 1, Table S5**). Positive cytoplasmic staining of ECHS1 was observed in prostate tumor tissue but not in BPH tissue (**Fig. 6**), although the difference among different cancer stages as indicated by Gleason scores (GS, from 6 to 9) was not significant. We also compared the RNAseq of normal and tumor in a TCGA dataset containing 549 patients (52 control patients and 497

tumor patients) [40]. ECSH1 also show significant different expression (**Fig. 6d**). This data confirms that ECHS1 is a promising prostate cancer biomarker.

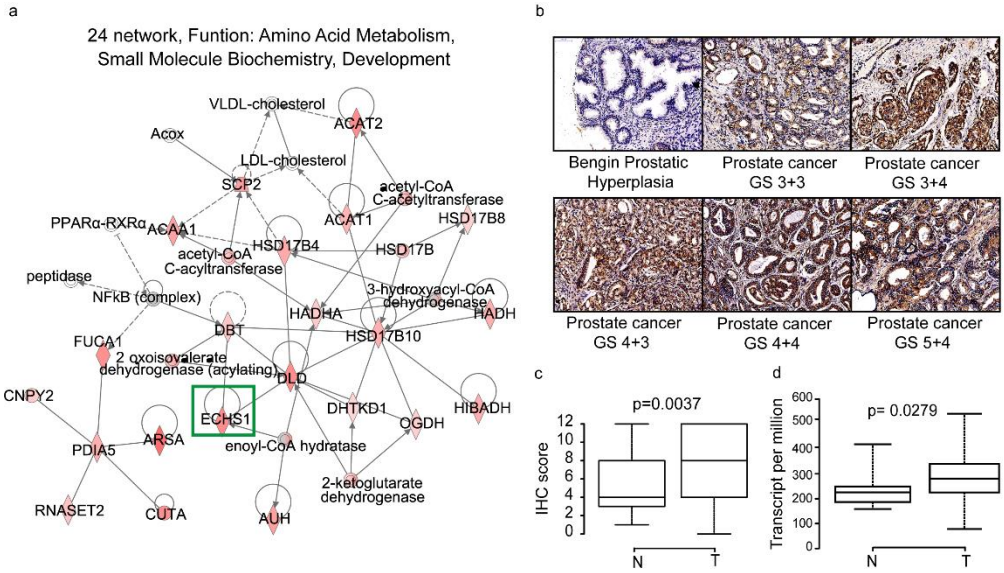

Figure 6. IHC stain for ECHS1 on tissue microarray. **(a)** IPA analysis shows Network No. 24 contains ECHS1. **(b)** TMA shows positive staining of ECHS1 in prostate cancer tissues. The staining images of representative BPH patients (patient SN: 201782041-3) and PCa patients with different Gleason Scores from 6 to 9 are shown (patient SN: GS3+3:201747411-7, GS3+4: 201738874-13, GS4+3: 201742345-10, GS4+4: 201754711-12, GS5+4: 201747749-17). **(c)** IHC score shows elevated expression of ECHS1 in prostate cancer tissues compared with BPH types. The p value was calculated using a student *t*-test. **(d)** Transcript copies were calculated from PCa and BPH tissue types based on the TCGA dataset. Student's *t*-test was used to calculate the significance.

Conclusion

The microflow S3 SWATH enables practical detection of regulated proteins in prostate cancer tissues which are largely identical to those identified by conventional SWATH method and consistent with targeted quantification using MRM^{HR} for shortlisted proteins. Proteins prioritized by the S3 SWATH method were further verified by two independent prostate cancer cohorts. Our multilayer data nominated ECHS1 as a promising biomarker for PCa and potential drug targets. This work presents a novel proteomics pipeline based on an accelerated microflow SWATH MS strategy with potential to accelerate the discovery and verification of protein biomarkers for precision medicine.

Author Contributions:

T.G., C.H., R.S. designed the project. C.H., N.M. and C.C. optimized the S3-SWATH. B.W., X.Y., X.T., L.C. procured the three prostate cohorts. R.S. performed the PCT-SWATH analysis with help from X.C. C.C. and R.S. performed the MRM^{HR} analysis. W.G., R.S., S.D., J.S. analyzed the data. R.S., Y.Z., C.H. and T.G. wrote the manuscript. C.Y., Z.L assisted data analysis. R.A. gave valuable advice. T.G., Y.Z. supported and supervised the project.

Research Funding: Zhejiang Provincial Natural Science Foundation of China (Grant No. LR19C050001 to T.G.). Hangzhou Agriculture and Society Advancement Program (Grant No. 20190101A04 to T.G.).

Acknowledgments: The authors thank all collaborators who participated in the procurement of the clinical specimens.

Competing financial interests: The research group of T.G. is partly supported by SCIEX, which provides access to prototype instrumentation, and Pressure Biosciences Inc, which provides access to advanced sample preparation instrumentation.

Data and materials availability: The S3 SWATH data are deposited in PRIDE. Project accession: IPX0001645000. The S3 SWATH data are deposited in iProX (IPX0001645001). The MRM^{HR} data are deposited in iProX (IPX0001645002). All the data will be publicly released upon publication.

References

1. Loeb, S., et al., *Overdiagnosis and overtreatment of prostate cancer.* Eur Urol, 2014. **65**(6): p. 1046-55.
2. Bache, N., et al., *A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics.* Mol Cell Proteomics, 2018. **17**(11): p. 2284-2296.
3. Bekker-Jensen, D.B., et al., *An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes.* Cell Syst, 2017. **4**(6): p. 587-599 e4.
4. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.* Mol Cell Proteomics, 2012. **11**(6): p. O111 016717.
5. Rost, H.L., et al., *OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data.* Nat Biotechnol, 2014. **32**(3): p. 219-23.
6. Guo, T., et al., *Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps.* Nat Med, 2015. **21**(4): p. 407-13.
7. Zhu, Y., et al., *Identification of Protein Abundance Changes in Hepatocellular Carcinoma Tissues Using PCT-SWATH.* Proteomics Clin Appl, 2019. **13**(1): p. e1700179.
8. Liu, Y., et al., *Quantitative variability of 342 plasma proteins in a human twin population.* Mol Syst Biol, 2015. **11**(1): p. 786.
9. Guo, T., et al., *Multi-region proteome analysis quantifies spatial heterogeneity of prostate tissue biomarkers.* Life Sci Alliance, 2018. **1**(2).
10. Shi, J., et al., *Comparison of protein expression between human livers and the hepatic cell lines HepG2, Hep3B, and Huh7 using SWATH and MRM-HR proteomics: Focusing on drug-metabolizing enzymes.* Drug Metab Pharmacokinet, 2018. **33**(2): p. 133-140.
11. He, B., et al., *Label-free absolute protein quantification with data-independent acquisition.* J Proteomics, 2019. **200**: p. 51-59.
12. Colgrave, M.L., et al., *Comparing Multiple Reaction Monitoring and Sequential Window Acquisition of All Theoretical Mass Spectra for the Relative Quantification of Barley Gluten in Selectively Bred Barley Lines.* Anal Chem, 2016. **88**(18): p. 9127-35.
13. Le Duff, M., et al., *Regulation of senescence escape by the cdk4-EZH2-AP2M1 pathway in response to chemotherapy.* Cell Death Dis, 2018. **9**(2): p. 199.
14. Vowinckel, J., et al., *Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition.* Sci Rep, 2018. **8**(1): p. 4346.
15. Liu, Y., et al., *Multi-omic measurements of heterogeneity in HeLa cells across laboratories.* Nat Biotechnol, 2019. **37**(3): p. 314-322.
16. Yi Zhu 1, 3*,Tobias Weiss 4*, Qiushi Zhang 1,2, Rui Sun 1,2, Bo Wang 5, Zhicheng Wu 1,2,

Qing Zhong 6,7, Xiao Yi 1,2 , Huanhuan Gao 1,2, Xue Cai 1,2, Guan Ruan 1,2, Tiansheng Zhu 1,2, Chao Xu, Sai Lou 9, Xiaoyan Yu 10, Ludovic Gillet 3, Peter Blattmann 3, Karim Saba 11, Christian D.Fankhauser 11, Michael B. Schmid 11, Dorothea Rutishauser 6, Jelena Ljubicic 6, Ailsa, Christiansen 6, Christine Fritz 6, Niels J. Rupp 6, Cedric Poyet 11, Elisabeth Rushing 12, Michael Weller 4, Patrick Roth 4, Eugenia Haralambieva 6, Silvia Hofer 13, Chen Chen 14, Wolfram Jochum 15, Xiaofei Gao 1,2, Xiaodong Teng 5, Lirong Chen 10, Peter J. Wild 6,16#, Ruedi Aebersold 3,17# , Tiannan Guo, *High-throughput proteomic analysis of FFPE tissue samples facilitates tumor stratification.* biorxiv, 2019.

17. Zhu, Y. and T. Guo, *High-Throughput Proteomic Analysis of Fresh-Frozen Biopsy Tissue Samples Using Pressure Cycling Technology Coupled with SWATH Mass Spectrometry.* Methods Mol Biol, 2018. **1788**: p. 279-287.
18. MacLean, B., et al., *Skyline: an open source document editor for creating and analyzing targeted proteomics experiments.* Bioinformatics, 2010. **26**(7): p. 966-8.
19. Rosenberger, G., et al., *A repository of assays to quantify 10,000 human proteins by SWATH-MS.* Sci Data, 2014. **1**: p. 140031.
20. Guo T1, Li L3, Zhong Q4,5, Rupp NJ4, Charmpi K3, Wong CE4, Wagner U4, Rueschoff JH4, Jochum W6, Fankhauser CD7, Saba K7, Poyet C7, Wild PJ4,8, Aebersold R1,9, Beyer A, *Multi-region proteome analysis quantifies spatial heterogeneity of prostate tissue biomarkers.* Life Sci Alliance., 2018.
21. Counihan, J.L., E.A. Grossman, and D.K. Nomura, *Cancer Metabolism: Current Understanding and Therapies.* Chem Rev, 2018. **118**(14): p. 6893-6923.
22. Wu, X., et al., *Lipid metabolism in prostate cancer.* Am J Clin Exp Urol, 2014. **2**(2): p. 111-20.
23. Thompson, T.C., et al., *Loss of p53 function leads to metastasis in ras+myc-initiated mouse prostate cancer.* Oncogene, 1995. **10**(5): p. 869-79.
24. Edwards, J., et al., *The role of c-Jun and c-Fos expression in androgen-independent prostate cancer.* J Pathol, 2004. **204**(2): p. 153-8.
25. Momparler, R.L., *Epigenetic therapy of cancer with 5-aza-2'-deoxycytidine (decitabine).* Semin Oncol, 2005. **32**(5): p. 443-51.
26. Panigrahy, D., et al., *PPARalpha agonist fenofibrate suppresses tumor growth through direct and indirect angiogenesis inhibition.* Proc Natl Acad Sci U S A, 2008. **105**(3): p. 985-90.
27. Saxman, S., et al., *Phase III trial of cyclophosphamide versus cyclophosphamide, doxorubicin, and methotrexate in hormone-refractory prostatic cancer. A Hoosier Oncology Group study.* Cancer, 1992. **70**(10): p. 2488-92.
28. Peterson, A.C., et al., *Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics.* Mol Cell Proteomics, 2012. **11**(11): p. 1475-88.
29. Lin, J.F., et al., *Identification of candidate prostate cancer biomarkers in prostate needle biopsy specimens using proteomic analysis.* Int J Cancer, 2007. **121**(12): p. 2596-605.
30. Whitaker, H.C., et al., *Peroxiredoxin-3 is overexpressed in prostate cancer and promotes cancer cell survival by protecting cells from oxidative stress.* Br J Cancer, 2013. **109**(4): p. 983-93.
31. Harter, C., et al., *Nonclathrin coat protein gamma, a subunit of coatomer, binds to the*

cytoplasmic dilysine motif of membrane proteins of the early secretory pathway. Proc Natl Acad Sci U S A, 1996. **93**(5): p. 1902-6.

32. Gerich, B., et al., *Non-clathrin-coat protein alpha is a conserved subunit of coatomer and in Saccharomyces cerevisiae is essential for growth.* Proc Natl Acad Sci U S A, 1995. **92**(8): p. 3229-33.
33. Iglesias-Gato, D., et al., *The Proteome of Primary Prostate Cancer.* Eur Urol, 2016. **69**(5): p. 942-52.
34. Zhang, Y.K., et al., *Enoyl-CoA hydratase-1 regulates mTOR signaling and apoptosis by sensing nutrients.* Nat Commun, 2017. **8**(1): p. 464.
35. Sharpe, A.J. and M. McKenzie, *Mitochondrial Fatty Acid Oxidation Disorders Associated with Short-Chain Enoyl-CoA Hydratase (ECHS1) Deficiency.* Cells, 2018. **7**(6).
36. Chang, Y., et al., *ECHS1 interacts with STAT3 and negatively regulates STAT3 signaling.* FEBS Lett, 2013. **587**(6): p. 607-13.
37. De Nunzio, C., et al., *The correlation between metabolic syndrome and prostatic diseases.* Eur Urol, 2012. **61**(3): p. 560-70.
38. Ylitalo, E.B., et al., *Subgroups of Castration-resistant Prostate Cancer Bone Metastases Defined Through an Inverse Relationship Between Androgen Receptor Activity and Immune Response.* Eur Urol, 2017. **71**(5): p. 776-787.
39. Twiddy, A.L., C.G. Leon, and K.M. Wasan, *Cholesterol as a potential target for castration-resistant prostate cancer.* Pharm Res, 2011. **28**(3): p. 423-37.
40. Chandrashekar, D.S., et al., *UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses.* Neoplasia, 2017. **19**(8): p. 649-658.