PulseDIA: in-depth data independent acquisition mass spectrometry using enhanced gas phase fractionation

Xue Cai [1,2], Weigang Ge [1,2], Xiao Yi [1,2], Rui Sun[1,2], Jiang Zhu [3], Cong Lu[3], Ping Sun[4], Tiansheng Zhu[1,2], Guan Ruan[1,2], Chunhui Yuan [1,2], Shuang Liang [1,2], Mengge Lyv [1,2], Shiang Huang [3], 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, China
2. Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou 310024, Zhejiang, China
3. Center for Stem Cell Research and Application, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, Hubei, China
4. Department of Hepatobiliary Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, Hubei, China

***** co-correspondence: zhuyi@westlake.edu.cn; guotiannan@westlake.edu.cn

ABSTRACT

An inherent bottleneck of data independent acquisition (DIA) analysis by Orbitrap-based mass spectrometers is the relatively large window width due to the relatively slow scanning rate compared to TOF. Here we present a novel gas phase separation and MS acquisition method called PulseDIA-MS, which improves the specificity and sensitivity of Orbitrap-based DIA analysis. This is achieved by dividing the ordinary DIA-MS analysis covering the entire mass range into multiple injections for DIA-MS analyses with complementary windows. Using standard HeLa digests, the PulseDIA method identified 69,530 peptide precursors from 9,337 protein groups with ten MS injections of 30 min LC gradient. The PulseDIA scheme containing two complementary windows led to the highest gain of peptide and protein identifications per time unit compared to the conventional 30 min DIA method. We further applied the method to profile the proteome of 18 cholangiocarcinoma (CCA) tissue samples (benign and malignant) from nine patients. PulseDIA identified 7,796 protein groups in these CCA samples, with 14% increase of protein identifications, compared to the conventional DIA method. The missing value for protein matrix dropped by 7% with PulseDIA acquisition. 681 proteins were significantly dysregulated in tumorous CCA samples. Together, we presented and benchmarked an alternative DIA method with higher sensitivity and lower missing rate.

INTRODUCTION

Mass spectrometry (MS)-based quantitative proteomics is increasingly applied to identify dysregulated proteins in clinical specimens, facilitating tumor diagnosis and prognosis.[1-3] DIA emerges recently as a significant discovery proteomics method enabling high-throughput and reproducible single-shot analysis of complex proteomes including those from clinical specimens.[3-6] DIA-MS acquires peptide precursors and fragment ion information using a predefined scanning scheme which is independent of the spectral data acquired, in contrast to data-dependent acquisition which selects peptide precursors for fragmentation based on their intensity values in the MS1 scan. The recent advance of DIA-MS, in particularly SWATH-MS[4], taking advantage of the fast scanning rate, fragments a packet of precursor ions after the MS1 scan in an incremental and repeating fashion (called DIA windows), then records the fragment signals of all of these peptide precursors in the respective MS2 scans for peptide identifications and quantifications.

Several DIA-MS methods have been reported, including all-ion fragmentation (AIF)[7] and MS^{E} [8]. In AIF and MS^{E}, all precursor ions are analyzed in one window, collecting MS1 and MS2 spectra alternatively. The thus obtained MS2 spectra are highly convoluted and could not be effectively analyzed with conventional algorithms for data-dependent acquisition (DDA) proteomics. The emerging DIA-MS reduces the spectral complexity by making use of gas-phase isolation of flying peptide precursors in multiple windows with certain mass-charge (*m/z*) width.[9] However, the sensitivity and specificity of DIA-MS are therefore limited by the window width. Heaven *et al* designed multiple DIA methods with three isolation sizes and various precursor ranges to systematically evaluate DIA for sensitive and reproducible proteomics[10]. Compared to 32 sequential DIA windows of 26 Daltons each across 400-1200 *m/z*, SWATH method with 56 sequential windows of 6 Daltons each across 450-730 *m/z* increases the signal-to-noise significantly.

When the number of isolation windows reaches the limit, gas-phase separation of peptide precursor ions could improve the sensitivity. A DIA-MS method called Precursor Acquisition Independent From Ion Count (PAcIFIC)[11-12] acquires tandem mass spectra with every 2.5 *m/z* segment for a 4.2-day nano-ESI method, leading to increased peptide and protein identifications. In 2016, Chapman *et al*[13] developed a 24-hour CSI PAcIFIC method which achieved minimal reduction of peptide and protein identifications with the PAcIFIC method above. The

throughput of PAcIFIC will likely further increase with cutting edge mass spectrometers. Another DIA method called multiplexing strategy (MSX) divides the peptide precursors in gas-phase into 100 windows.[14] To demultiplex the spectra, five separate 4-*m/z* isolation windows are analyzed in each scan of MSX DIA, and 20 scans are performed to cover the *m/z* range of 500-900. Both precursor ion selectivity (5-fold higher than the conventional DIA) and fragment-ion spectra quality are improved under the narrower window width. In addition to the gas-phase separation, overlapping windows were designed for a DIA strategy which improves precursor selectivity without any loss in other key acquisition parameters.[15]

Here, we present an alternative gas-phase separation PulseDIA-MS method, in which replicate sample injections are analyzed using differentially segmented DIA-MS methods in order to decrease window width and improve specificity and sensitivity. We demonstrated its applicability in identifying dysregulated proteins from cholangiocarcinoma (CCA), a rare malignant tumor composed of cells that resemble those of the biliary tract.[16] CCA usually eludes from early diagnosis due to hardly discernable symptoms, leading to a poor prognosis and high morbidity. The serum marker carbohydrate antigen 19-9 (CA19-9) and carcinoembryogenic antigen (CEA) are sometimes used to detect CCA, however they are neither sensitive nor specific.[17] The need for discovering proteomic biomarkers for early diagnosis of CCA is pressing. Several proteomic studies of cell lines, bile fluid and sera have been reported.[18-20] Juliet P *et al*. used DDA-MS to identify several promising biomarkers (such as ANXA1, ANXA10, ANXA13) of CCA by proteomic analysis of microdissected cells extracted from 11 tissue samples collected from 11 CCA patients and verification by immunohistochemistry (IHC) of other 83 samples.[21]

In this study, we performed proteomic analysis of 18 CCA tissue samples by the thus developed PulseDIA-MS method. 7,796 protein groups were quantified across all CCA samples and 681 proteins were found to be dysregulated significantly.

EXPERIMENTAL SECTION

Samples

HeLa Protein Digest Standard peptides were purchased from the Thermo Fisher Scientific™ (Product number 88329, Rockford, USA), and stored at -20□ until analysis. All cell lines were provided by Dr Chenhuan Yu from Zhejiang Academy of Medical Sciences, Hangzhou, China. BT549, Hs578T, ZR75-1 and MDA-MB-231 cells were cultured in DMEM/F-12 medium (Cat No. 01-172-1ACS, Biological industries, Cromwell, CT, USA). MDA-MB-468 was maintained in L15 medium (Cat No. 01-115-1A, Biological industries, Cromwell, CT, USA). T47D was adapted in the DMEM medium (Cat No. 06-1055-57-1ACS, HyClone, Biological industries, Cromwell, CT, USA). MCF7, MX-1 and SK-BR-3 cells were cultured in 1640 medium (Cat No. 01-100-1ACS, Biological industries, Cromwell, CT, USA) supplemented with 10% fetal bovine serum (Cat No. 04-001AUS-1A, Biological industries, Cromwell, CT, USA) and penicillin-streptomycin (Cat No. 03-031-5B, HyClone, General Electric, USA) at 37 □ with 5% CO_2.

18 tissue samples from nine CCA patients were collected within one hour after hepatectomy, then snap frozen and stored at -80□. A pair of tumorous tissue and the non-tumorous tissue from an adjacent region around the tumor as determined by histomorphological examination were collected from the same patient. Ethical permission was approved by Ethics Committee of Tongji Medical College, Huazhong University of Science and Technology. This study was also approved by Ethics Committee of Westlake University.

Peptide preparation

HeLa Protein Digest Standard peptides were re-dissolved with HPLC-grade water containing 0.1% formic acid (FA) and 2% acetonitrile (ACN) at the final concentration of 0.25 μg/μL. Cell pellets and tissue samples were lysed and digested using pressure cycling technology (PCT) as described previously with some modifications.[22] Briefly, the cells were lysed with 50 μL lysis buffer containing 6 M urea (Sigma) and 2 M thiourea (Sigma) in 100 mM ammonium bicarbonate in a PCT -MicroTube. The PCT-assisted lysis was performed under the scheme with 60 cycles with each cycle consisting of 30 s at 45,000 p.s.i. and 10 s at ambient pressure at 30 □. PCT-assisted digestion was performed using lysC (1:40) and trypsin (1:50) in a barocycler under the PCT scheme for 45 cycles and 60 cycles, respectively. Each cycle contains 50 s at 20,000 p.s.i. and 10 s at ambient pressure at 30 □. The thus generated peptides were acidified with trifluoroacetic acid (TFA) to pH 2-3, and cleaned with the Nest Group C18 columns (17-170 μg capacity, Part No. HEM S18V, MA, USA) prior to MS analysis. Nine breast cancer cell lines were prepared separately for building DIA spectral library and a pooled sample for nine breast cancer cell lines were prepared for both building DIA spectral library and PulseDIA optimization.

Data acquisition with mass spectrometry

With the PulseDIA method, the MS1 was performed over a *m/z* range of 390-1210 with resolution at 60,000,

AGC target of 3e6, and maximum ion injection time of 80 ms. The isolation windows for PulseDIA method were complementary for the same set of pulse injections. For instance, if the injection number of PulseDIA is four, there are four complementary isolation windows which corresponding to four injections (**Figure 1**). All the isolation windows for this study are provided in the supplementary **Table S1-4**. The MS2 was performed with resolution at 30,000, AGC target of 1e6, and maximum ion injection time of 50 ms.

The PulseDIA acquisition of HeLa peptides was performed on a nanoflow EASY-nLC™ 1200 System (Thermo Fisher Scientific™, San Jose, USA) coupled to a Q Exactive HF-X hybrid Quadrupole-Orbitrap (Thermo Fisher Scientific™, San Jose, USA). For each PulseDIA acquisition, 0.5 μg of peptides was injected and separated across a 30 min LC gradient (from 8% to 40% buffer B) at a flowrate of 300 nl/min (precolumn, 3 μm, 100 Å, 20 mm*75 μm i.d.; analytical column, 1.9 μm, 120 Å, 150 mm*75 μm i.d.). Buffer A was HPLC-grade water containing 0.1% FA, and buffer B was 80%ACN, 20%H_2O containing 0.1%FA.

The PulseDIA acquisition of peptides from other cell lines, CCA tissues and mixtures of HeLa and *E .coli* digests were performed on a nanoflow DIONEX UltiMate 3000 RSLCnano System (Thermo Fisher Scientific™, San Jose, USA) coupled to a Q Exactive HF hybrid Quadrupole-Orbitrap (Thermo Fisher Scientific™, San Jose, USA). For each PulseDIA acquisition, except for the mixtures of HeLa and *E. coli* digests, 0.5 μg of peptides was injected and separated across a 30 min LC gradient with the same settings as described above.

With the DDA method, the MS1 was performed over a *m/z* range of 400-1200 with the resolution at 60,000, AGC target of 3e6, and maximum ion injection time of 80 ms. The MS2 was performed for top 20 precursors with resolution at 30,000, AGC target of 1e5, and maximum ion injection time of 100 ms.

The DDA acquisition of peptides from cell lines was performed on a nanoflow a nanoflow EASY-nLC™ 1200 System (Thermo Fisher Scientific™, San Jose, USA) coupled to a Q Exactive HF-X hybrid Quadrupole-Orbitrap (Thermo Fisher Scientific™, San Jose, USA). For each DDA acquisition, 0.5 μg of peptides was injected and separated over a 90 min LC gradient with the same settings as described above.

PulseDIA data analysis

PulseDIA raw files were converted into mzML format using msconvert[23] and analyzed using DIA-NN (1.6.0)[24] against suitable spectral library[25]. The first library named BRP DIA library (provided in **Data deposition**) containing 63,189 peptide precursors and 6,043 proteins was built by Spectronaut[26] software with default settings for the breast cancer cell line samples analysis. Fifteen DDA files containing nine files from each cell line and six files from the pooled sample were used to build BRP DIA library. The second library named HeLa DIA spectral library (provided in **Data deposition**) containing 291,236 peptide precursors and 17,338 protein groups. The third library named HeLa *E.coli* DIA library (provided in **Data deposition**) containing 15,455 *E.coli* peptide precursors, 59,206 HeLa peptide precursors and 1,920 *E.coli* proteins and 6,536 HeLa proteins, was built by Spectronaut software for HeLa *E.coli* spike-in samples analysis. The last library named DIA Pan Human Library (DPHL) containing 396,245 peptide precursors and 14,786 protein groups (manuscript in press) was used for CCA tissue samples analysis.

In the DIA-NN setting, the peptide length range was set from 7 to 30, precursor *m/z* range was set from 400 to 1200, and fragment ion *m/z* range was set from 100 to 1500. The mass accuracy and chromatogram scan window size were set by the software automatically. Peptide precursors and protein FDRs were controlled below 1%. The quantitative result of proteins in multiple injections (excluding null values) were extracted from the result of DIA-NN using a R program named PulseDIA_DIANNreport_extract (https://github.com/Allen188/PulseDIA), and then combined into the protein combined matrix using a R program named PulseDIA_DIANNreport_combine (https://github.com/Allen188/PulseDIA). We took the average value as the final quantitative results for the proteins identified by different injections of PulseDIA or Duplicate PulseDIA.

Statistical analysis

Pearson correlation coefficient (r) was calculated using the cor function in R with the "pairwise.complete.obs" setting on based on the commonly identified peptide precursors and proteins. P value was calculated using two-tailed, paired student's t test, and further adjusted using Benjamini-Hochberg (BH) method.

RESULTS AND DISCUSSION

Design of the PulseDIA Acquisition

Most peptide precursors in a proteome locate in the 400-1200 *m/z* mass space. Generally shorter and hydrophilic peptides are of lower *m/z* while longer and hydrophobic peptides have higher *m/z* values. The density of peptide precursors in the mass range of 400-800 is higher than the rest. Therefore, conventionally the 400-800 *m/z* range is divided into 20 windows (each 20 Thomas), while large windows (60-140 Thomas) are schemed in the

800-1200 *m/z* range (Figure 1). In DIA coupled with targeted data analysis, smaller windows usually lead to higher sensitivity and specificity, however, the window size is limited by the MS scanning rate. In PulseDIA, we utilized multiple injections to achieve small DIA windows for better gas phase separation of the peptide precursors. Each DIA window in the 400-1200 *m/z* range is divided into n times smaller range of *m/z* by evenly dividing it into n MS injections in a pulse manner. For instance, if n is four, the PulseDIA evenly divides each window from the classical setting to four portions, and then allocates them into four MS injections sequentially. As a result, the PulseDIA acquires data for 96 windows with 1/4 mass range width of the original ones. We also compiled a duplicate PulseDIA method in which each window doubles its range compared to the standard PulseDIA method and have 50% mass range overlap with its two adjacent windows. The rationale is that with this scheme, peptide precursors in each window range will be fragmented twice and their data are acquired in two independent injections. This design may improve the quantitative accuracy and reproducibility. Multiple injections require higher amount of samples, however, this is normally not a big issue. With PCT-assisted sample preparation, we usually generate 50 μg peptides from 1 mg tissue samples[3, 27] which are sufficient for about 100 injections.

Customized PulseDIA window scheme can be generated using a R program named PulseDIA_calcu_wins (https://github.com/Allen188/PulseDIA). Parameters including MS1 acquisition range, number of windows, number of injection fractions, whether overlap of windows is allowed, and fixed or variable window scheme can be defined in the script for PulseDIA window scheme generation. As to variable windows, the precursor ions density information is required. With the four-pulse scheme, one would get four PulseDIA raw data for one sample. All the isolation window tables in this experiment were provided in the supplementary **Table S1-4**.

The PulseDIA method is different from the published PAcIFIC[11-12]. PAcIFIC divides precursors into windows as narrow as 2.5 Thomas with overlaps so that the data can be analyzed with shotgun proteomics search engines such as Sequest against protein sequence databases[13], which assumes that relatively pure peptide precursors are isolated in gas phase and produce relatively pure MS2 fragment ions for peptide sequence inference. Each MS2 spectrum is analyzed individually. Whereas in PulseDIA, the *m/z* windows are of a larger range (min 10 Thomas, max 70 Thomas, mean ~20 Thomas with two pulse injections), and data are analyzed in a targeted manner. In the PulseDIA data analysis, no MS2 spectrum is analyzed in isolated fashion, instead multiple MS2 spectra along the retention time are subject to extracted ion chromatographic analysis for peak group construction which are further scored to identify and quantify peptide precursors. Therefore, the PulseDIA window size is more flexible, and does not require isolation of a particular peptide precursor ions.

The PulseDIA is also different from conventional DIA method and DIA with gas-phase fractionation using consecutively separated windows[15, 28] coupled with targeted data analysis by acknowledging the fact that peptide precursor ions are not evenly distributed long the *m/z* space. The PulseDIA utilizes discontinuous windows which allocate peptide precursor ions with different *m/z* evenly into each pulse injection. In this way, each injection contains similar number of peptide precursors with diverse length, *m/z*, charge state and hydrophobicity, which greatly facilitates targeted data analysis and retention time alignment when multiple samples are analyzed.

To analyze the PulseDIA data (**Figure S1**), all the PulseDIA raw data were converted to mzML format using msconvert software for DIA-NN analysis.

Optimization of PulseDIA

A mixed peptide digest sample from nine breast cancer cell lines was firstly used to evaluate the technical reproducibility of the method, and then to optimize PulseDIA parameters systematically. Results showed that proteins were identified at a high degree of reproducibility. The coefficients of variation (CV) values were calculated for each pair of technical replicates. While missing values were excluded from the CV calculation, an average of 93% quantified proteins were identified in technical replicates, therefore the influence of missing value on the CV values is not substantial. As shown in **Figure S2**, the median CV values are around 1.0 - 2.0%. Four parameters were tested and optimized to maximize the performance of PulseDIA: i) number of injections; ii) length of LC gradient; iii) fixed or variable window; iv) PulseDIA or duplicate PulseDIA. As shown in Figure S3, more peptide precursors and proteins were identified with more injections and longer LC gradient. PulseDIA with fixed windows led to comparable number peptide precursor and protein identifications than that with variable windows (Figure S3). PulseDIA is slightly better than duplicate PulseDIA in terms of peptide precursor and protein identification, particularly for short gradient, probably due to doubled isolation window width in duplicate PulseDIA (Figure S3).

The highest number of IDs were achieved with five-injection PulseDIA scheme. A total of 42,563 peptide precursors and 5,223 protein groups were identified, covering 67% of peptide precursors and 86% protein groups of the library (**Table S5**). With a more comprehensive breast cancer cell line library, the ID will likely further increase. We then computed the gain of protein ID with increase LC gradient time, and found that the number of

increased proteins identified per increased LC gradient time unit (min) reach maximum with two PulseDIA runs of 30 min LC gradient compared to the conventional 30 min DIA analysis (**Figure S4**). We further compared the PulseDIA with two injections of 30 min LC gradient with the conventional DIA run of 60 min LC gradient (**Figure S5**), and found that the peptide precursor and protein identifications increased by 13.5% and 6.7% on average, respectively. It's worth noting that we computed only the effective LC elution gradient time, excluding sample injection, column wash and equilibrium time.

Next, we used the HeLa Protein Digest Standard peptides, since we have a more comprehensive spectral library, to verify the prioritized parameters of PulseDIA which were obtained from the breast cancer cell line samples. In order to test the PulseDIA capacity for maximum peptide and protein identification, a 10-injection PulseDIA runs of 30 min LC gradient were applied. As shown in **Figure 2a**, 69,530 peptide precursors and 9,337 protein groups were identified under this 10-injection PulseDIA scheme (**Table S6**) against the HeLa DIA spectral library. Compared to the 30 min LC gradient of conventional DIA, the peptide precursor and protein identifications increased 81.8 % and 38.5%, respectively. Besides, with the increase of injection number, the number of peptide precursor and protein identifications both increased too (**Figure 2a**). Similarly, two PulseDIA runs of 30 min LC gradient led to the maximum increase of peptide precursor and protein IDs per time unit compare to the conventional DIA with 30 min LC and 24 standard DIA windows (**Figure 2b**).

Next, we evaluated the quantitative accuracy of PulseDIA and the conventional DIA using several mixtures of HeLa and *E. coli* digests following the comparison performed by Heaven *et al*[13] (**Table S6**). The ROC plot in **Figure 2c** shows that PulseDIA has a higher quantitative accuracy than DIA. Each pair of technical replicates shared over 85% peptide precursors and proteins of the peptide precursors and proteins quantified in each sample. For the overlapped peptide precursors and proteins, the r values between two technical replicates are all greater than 0.91, indicating high repeatability and stability (**Figure 2d**).

Application of PulseDIA to proteotyping of CCA

After benchmarking the PulseDIA method with cell line samples, we next evaluated its applicability to clinical tissue samples. Both PulseDIA-MS and the conventional DIA-MS were applied to the proteomics profiling of 18 tissue samples (benign and tumor pairs) from nine CCA patients. The relevant clinical patient information and MS raw data were listed in supplementary **Table S7a and Table S7b**. These CCA peptide samples were analyzed with two MS strategies, i.e., the PulseDIA scheme containing two MS runs of 30 min LC gradient, and the conventional DIA-MS of 60 min LC gradient.

Results showed that PulseDIA identified more peptide precursors and proteins than conventional DIA in each sample using the same elution gradient excluding the sample injection, column wash and equilibration time (**Figure S6**). The increased percentage of peptide precursors and protein identification reach up to 56% and 47% respectively (**Table S7c**). In detail, for 18 tissue samples, PulseDIA identified 83,972 peptide precursors and 7,796 protein groups from 36 PulseDIA injections (two injections per sample), improved by 16% and 14% compared to DIA from 18 DIA injections (one injection per sample), respectively. We further checked the peptide and protein identification in different tissues (bebign vs. tumor). In peptides from tumorous tissues, PulseDIA identified a total of 75,907 peptide precursors and 7,482 protein groups, while conventional DIA identified 63,040 peptide precursors and 6,460 protein groups, both against the DPHL library **(Figure 3a**). As to tissue from the adjacent benign area, a total of 63,143 peptide precursors and 6,568 protein groups were identified by PulseDIA, while a total of 53,562 peptide precursors and 5,607 protein groups were identified using conventional DIA.

Proteomic data sets contain missing values which are proteins identified and quantified in some samples but not in some others in a data set. Missing values are sometimes due to technical issues, and multiple methods have been employed to impute these missing values for data analysis[29-30]. A recent paper reported that missing values in SWATH data are also biological.[31] The percentage of missing values, *i.e.* proteins not identified in some samples but in others, in the protein matrix generated by PulseDIA (35%) is lower than that generated by conventional DIA (42%) (**Figure 3b**). The PulseDIA and DIA analysis shared 83% peptide precursors and 91% proteins. The r values between two technical replicates for peptide precursors and proteins were 0.96 and 0.97 for PulseDIA while 0.83 and 0.86 for DIA, respectively, indicating that PulseDIA achieved better reproducibility than DIA probably due to increased specificity (**Figure 3c**). The proteins quantified by DIA and PulseDIA showed high consistency (r =0.91) (**Figure 3d**). Moreover, 681 significantly dysregulated proteins (adjusted p value <0.05, $|\log2(FC)| > 1$) were identified between tumorous and benign tissues by PulseDIA, while 434 proteins were identified by DIA (**Figure 3e**).

We then analyzed the 681 and 434 significantly regulated proteins using Ingenuity Pathway Analysis (IPA) to identify enriched pathways, respectively (**Table S8a, S8b**).[32] The data showed that 22 of top 25 enriched pathways were identical for both methods (**Table S8c, S8d**). In addition, the analysis of the upstream regulators also

yielded similar results, with 22 of top 25 upstream regulators were identical (**Table S8e, S8f**). PulseDIA led to identification of 25 networks while DIA of 21(**Table S8g, S8h**). The PulseDIA identified several dysregulated proteins which were absent in the DIA analysis, including NQO1, AHCY, MCC and MYEF2. NQO1 have been reported over-expressed in CCA [33], and AHCY was associated with adult onset hepatocellular carcinoma[34] , supporting our PulseDIA-based analysis.

CONCLUSION

In this study, we developed and benchmarked an alternative DIA-MS acquisition method called PulseDIA. Compared to the conventional DIA method, the PulseDIA method demonstrated higher sensitivity and specificity in peptide precursor and protein identification. A total of 69,530 peptide precursors and 9,337 protein groups were identified using ten PulseDIA runs of HeLa digest with 30 min LC gradient, which achieved an increase of 81.8% and 38.5% at peptide precursor and protein group level, respectively, compared to the conventional 30 min gradient DIA. The reproducibility of PulseDIA is high ($r>0.9$). We further applied PulseDIA to analyze biopsy-level CCA tissue samples and identified 83,972 peptide precursors and 7,796 protein groups. Compared to the conventional DIA method, the PulseDIA identified and 14% more proteins and reduced the missing value rate by 7%. We identified 681 significantly regulated proteins in CCA samples, uncovering novel protein biomarker candidates for CCA. Our case study showed that the PulseDIA method can be practically applied to protein biomarker research using clinical specimens with higher specificity, sensitivity and reproducibility.

Multiple injections increase the time for sample injection, column washing and equilibrium, which is non-trivial in ordinary nano-LC system. Adoption of two-trap column nano-LC system, microflow system and preformed gradients using Evosep[35] will further increase the throughput of PulseDIA.

ASSOCIATED CONTENT

Supporting Information

Workflow for PulseDIA data analysis, optimization result for breast cancer lines, number of peptide and protein identification for each CCA tissue sample (pdf). All isolation windows for MS method in this study (xlsx). Protein quantitative results of breast cancer lines (xlsx), HeLa digest (xlsx). Sample information, protein quantitative result, IPA analysis result of all CCA tissue samples(xlsx).

Data deposition

All the raw data and spectral library files in this report have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the iProX partner repository[36] with the dataset identifier PXD016904. Project accession of the breast cancer lines data: IPX0001769001. Project accession of the HeLa data: IPX0001769002. Project accession of the CCA data: IPX0001769003. All the data will be publicly released upon publication.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (General Program) (Grant No. 81972492 to T.G.), National Science Fund for Young Scholars (Grant No. 21904107), Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars (Grant No. LR19C050001 to T.G.), Hangzhou Agriculture and Society Advancement Program (Grant No. 20190101A04 to T.G.). We thank Chenhuan Yu for providing breast cancer cell lines.

AUTHOR CONTRIBUTIONS

T.G., X.C., Y.Z. designed the project. J.Z., C.L., P.S. procured the CAA cohorts. X.Y., R.S. performed the PCT-based sample preparation. X.C performed the PulseDIA analysis. W.G., X.C. T.Z analyzed the data. W.G., X.C., G.R. drew the graphs. X.C., T.G., Y.Z., C.Y., S.L., S.H., M.L. wrote the manuscript with inputs from all co-authors. T.G. supported and supervised the project.

CONFLICT OF INTEREST

The research group of T.G. is supported by Pressure Biosciences Inc, which provides access to advanced sample preparation instrumentation.

REFERENCES

1. Aebersold, R.; Mann, M., Mass-spectrometric exploration of proteome structure and function. *Nature.* **2016,** *537* (7620), 347-355.
2. Gillet, L. C.; Leitner, A.; Aebersold, R., Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. *Annual review of analytical chemistry.* **2016,** *9*, 449-472.
3. Guo, T.; Kouvonen, P.; Koh, C. C.; Gillet, L. C.; Wolski, W. E.; Rost, H. L.; Rosenberger, G.; Collins, B. C.; Blum, L. C.; Gillessen, S.; Joerger, M.; Jochum, W.; Aebersold, R., Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat Med.* **2015,** *21* (4), 407-13.
4. Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. J. M.; Proteomics, C., Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. **2012,** *11* (6), O111. 016717.
5. Liu, Y.; Buil, A.; Collins, B. C.; Gillet, L. C.; Blum, L. C.; Cheng, L. Y.; Vitek, O.; Mouritsen, J.; Lachance, G.; Spector, T. D., Quantitative variability of 342 plasma proteins in a human twin population. *Molecular Systems Biology.* **2015,** *11* (2), 786.
6. Venable, J. D.; Dong, M.-Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R., Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods.* **2004,** *1* (1), 39-45.
7. Geiger, T.; Cox, J.; Mann, M., Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Molecular & cellular proteomics : MCP.* **2010,** *9* (10), 2252-2261.
8. Plumb, R. S.; Johnson, K. A.; Paul, R.; Smith, B. W.; Wilson, I. D.; Castro-Perez, J. M.; Nicholson, J. K., UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun Mass Spectrom.* **2006,** *20* (13), 1989–1994.
9. Heaven, M. R.; L, C. A.; Yuan-Wei, N.; B., G. D.; W, H. A.; P, G. H.; M., C. R.; L., N. J., microDIA (μDIA): data-independent acquisition for high-throughput proteomics and sensitive peptide mass spectrum identification. *Analytical Chemistry.* acs.analchem.8b01026-.
10. Heaven, M. R.; Funk, A. J.; Cobbs, A. L.; Haffey, W. D.; Norris, J. L.; McCullumsmith, R. E.; Greis, K. D., Systematic evaluation of data-independent acquisition for sensitive and reproducible proteomics-a prototype design for a single injection assay. *Journal of mass spectrometry : JMS.* **2016,** *51* (1), 1-11.
11. Panchaud, A.; Scherl, A.; Shaffer, S. A.; von Haller, P. D.; Kulasekara, H. D.; Miller, S. I.; Goodlett, D. R., Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal Chem.* **2009,** *81* (15), 6481-8.
12. Panchaud, A.; Jung, S.; Shaffer, S. A.; Aitchison, J. D.; Goodlett, D. R., Faster, quantitative, and accurate precursor acquisition independent from ion count. *Anal Chem.* **2011,** *83* (6), 2250-7.
13. Chapman, J. D.; Edgar, J. S.; Goodlett, D. R.; Goo, Y. A., Use of captive spray ionization to increase throughput of the data-independent acquisition technique PAcIFIC. *Rapid Commun Mass Spectrom.* **2016,** *30* (9), 1101-7.
14. Egertson, J. D.; Kuehn, A.; Merrihew, G. E.; Bateman, N. W.; MacLean, B. X.; Ting, Y. S.; Canterbury, J. D.; Marsh, D. M.; Kellmann, M.; Zabrouskov, V.; Wu, C. C.; MacCoss, M. J., Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods.* **2013,** *10* (8), 744-6.
15. Amodei, D.; Egertson, J.; MacLean, B. X.; Johnson, R.; Merrihew, G. E.; Keller, A.; Marsh, D.; Vitek, O.; Mallick, P.; MacCoss, M. J., Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. *J Am Soc Mass Spectrom.* **2019,** *30* (4), 669-684.
16. Shimoda, M.; Kubota, K., Multi-disciplinary treatment for cholangiocellular carcinoma. *World J Gastroenterol.* **2007,** *13* (10), 1500-1504.
17. Silsirivanit, A.; Sawanyawisuth, K.; Riggins, G. J.; Wongkham, C., Cancer biomarker discovery for cholangiocarcinoma: the high-throughput approaches. *J Hepatobiliary Pancreat Sci.* **2014,** *21* (6), 388-396.
18. Scarlett, C. J.; Saxby, A. J.; Nielsen, A.; Bell, C.; Samra, J. S.; Hugh, T.; Baxter, R. C.; Smith, R. C., Proteomic profiling of cholangiocarcinoma: Diagnostic potential of SELDI-TOF MS in malignant bile duct stricture. *Hepatology.* **2006,** *44* (3), 658-666.
19. Lankisch, T. O.; Metzger, J.; Negm, A. A.; Voβkuhl, K.; Schiffer, E.; Siwy, J.; Weismüller, T. J.; Schneider, A. S.; Thedieck, K.; Baumeister, R.; Zürbig, P.; Weissinger, E. M.; Manns, M. P.; Mischak, H.; Wedemeyer, J., Bile proteomic profiles differentiate cholangiocarcinoma from primary sclerosing cholangitis and choledocholithiasis. *Hepatology.* **2011,** *53* (3), 875-884.
20. Mustafa, M. Z.; Nguyen, V. H.; Le Naour, F.; De Martin, E.; Beleoken, E.; Guettier, C.; Johanet, C.; Samuel, D.; Duclos-Vallee, J.-C.; Ballot, E., Autoantibody signatures defined by serological proteome analysis in sera from patients with cholangiocarcinoma. *J Transl Med.* **2016,** *14*, 17-17.
21. Padden, J.; Ahrens, M.; Kälsch, J.; Bertram, S.; Megger, D.; Bracht, T.; Eisenacher, M.; Kocabayoglu, P.; Meyer, H.; Sipos, B.; Baba, H.; Sitek, B., Immunohistochemical Markers Distinguishing Cholangiocellular Carcinoma from Pancreatic Ductal Adenocarcinoma Discovered by Proteomic Analysis of Microdissected Cells. *Molecular & cellular proteomics : MCP.* **2015,** *15*.
22. Zhu, Y.; Guo, T., High-Throughput Proteomic Analysis of Fresh-Frozen Biopsy Tissue Samples Using Pressure Cycling Technology Coupled with SWATH Mass Spectrometry. In *Tissue Proteomics: Methods and Protocols*, Sarwal, M. M.; Sigdel, T. K., Eds. Springer New York: New York, NY, 2018; pp 279-287.
23. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P., ProteoWizard: open source software for rapid

proteomics tools development. *Bioinformatics.* **2008,** *24* (21), 2534-2536.
24. Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M., DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods.* **2019**.
25. Schubert, O. T.; Gillet, L. C.; Collins, B. C.; Navarro, P.; Rosenberger, G.; Wolski, W. E.; Lam, H.; Amodei, D.; Mallick, P.; MacLean, B.; Aebersold, R., Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols*. **2015,** *10*, 426.
26. Bernhardt, O. M.; Selevsek, N.; Gillet, L. C.; Rinner, O.; Picotti, P.; Aebersold, R.; Reiter, L. J. B. c., Spectronaut: A fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data. **2012**.
27. Zhu, Y.; Weiss, T.; Zhang, Q.; Sun, R.; Wang, B.; Yi, X.; Wu, Z.; Gao, H.; Cai, X.; Ruan, G.; Zhu, T.; Xu, C.; Lou, S.; Yu, X.; Gillet, L.; Blattmann, P.; Saba, K.; Fankhauser, C. D.; Schmid, M. B.; Rutishauser, D.; Ljubicic, J.; Christiansen, A.; Fritz, C.; Rupp, N. J.; Poyet, C.; Rushing, E.; Weller, M.; Roth, P.; Haralambieva, E.; Hofer, S.; Chen, C.; Jochum, W.; Gao, X.; Teng, X.; Chen, L.; Zhong, Q.; Wild, P. J.; Aebersold, R.; Guo, T., High-throughput proteomic analysis of FFPE tissue samples facilitates tumor stratification. *Molecular Oncology.* **2019,** *13* (11), 2305-2328.
28. Ting, Y. S.; Egertson, J. D.; Bollinger, J. G.; Searle, B. C.; Payne, S. H.; Noble, W. S.; MacCoss, M. J., PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nature methods.* **2017,** *14* (9), 903-908.
29. Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T., Accounting for the multiple natures of missing values in label-free quantitative proteomics datasets to compare imputation strategies. *Journal of Proteome Research.* acs.jproteome.5b00981.
30. Forstenlehner, I. C.; Holzmann, J.; Toll, H.; Huber, C. G., Site-specific characterization and absolute quantification of pegfilgrastim oxidation by top-down high-performance liquid chromatography-mass spectrometry. *Anal Chem.* **2015,** *87* (18), 9336-43.
31. McGurk, K. A.; Dagliati, A.; Chiasserini, D.; Lee, D.; Plant, D.; Baricevic-Jones, I.; Kelsall, J.; Eineman, R.; Reed, R.; Geary, B.; Unwin, R. D.; Nicolaou, A.; Keavney, B. D.; Barton, A.; Whetton, A. D.; Geifman, N., The use of missing values in proteomic data-independent acquisition mass spectrometry to enable disease activity discrimination. *Bioinformatics.* **2019**.
32. Andreas, K. M.; Jeff, G.; Jack, P.; Stuart, T., Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics.* **2014,** *30* (4), 523.
33. Buranrat, B.; Chau-In, S.; Prawan, A.; Puapairoj, A.; Kukongviriyapan, V., NQO1 Expression Correlates with Cholangiocarcinoma Prognosis. *Asian Pac J Cancer Prev.* **2012,** *13* (5), 131-136.
34. Belužić, L.; Grbeša, I.; Belužić, R.; Park, J. H.; Kong, H. K.; Kopjar, N.; Espadas, G.; Sabidó, E.; Lepur, A.; Rokić, F.; Jerić, I.; Brkljačić, L.; Vugrek, O., Knock-down of AHCY and depletion of adenosine induces DNA damage and cell cycle arrest. *Scientific Reports.* **2018,** *8* (1), 14012.
35. Bache, N.; Geyer, P. E.; Bekker-Jensen, D. B.; Hoerning, O.; Falkenby, L.; Treit, P. V.; Doll, S.; Paron, I.; Müller, J. B.; Meier, F.; Olsen, J. V.; Vorm, O.; Mann, M., A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. *Molecular & cellular proteomics : MCP.* **2018,** *17* (11), 2284-2296.
36. Jie; Ma; Tao; Chen; Songfeng; Wu; Chunyuan; Yang; Mingze; Bai, iProX: an integrated proteome resource.

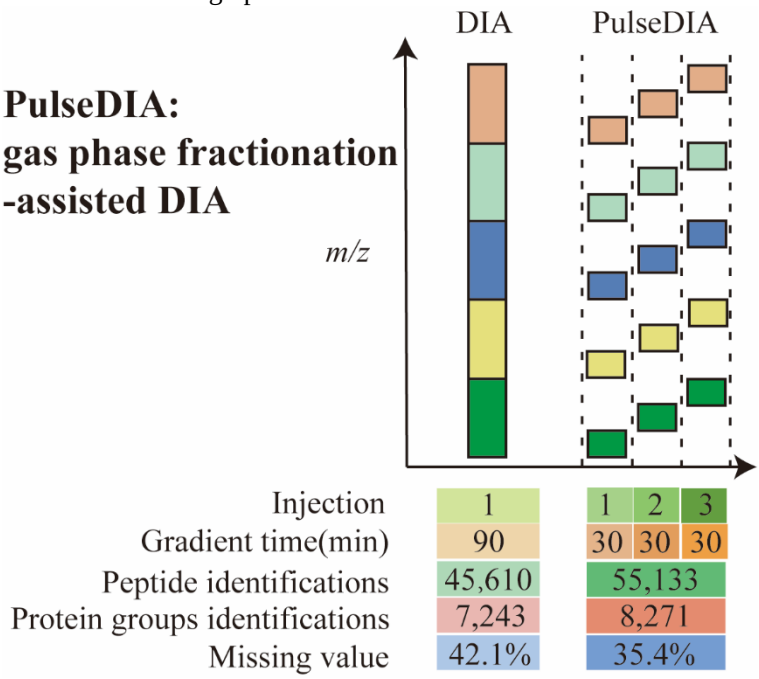
Figure captions：

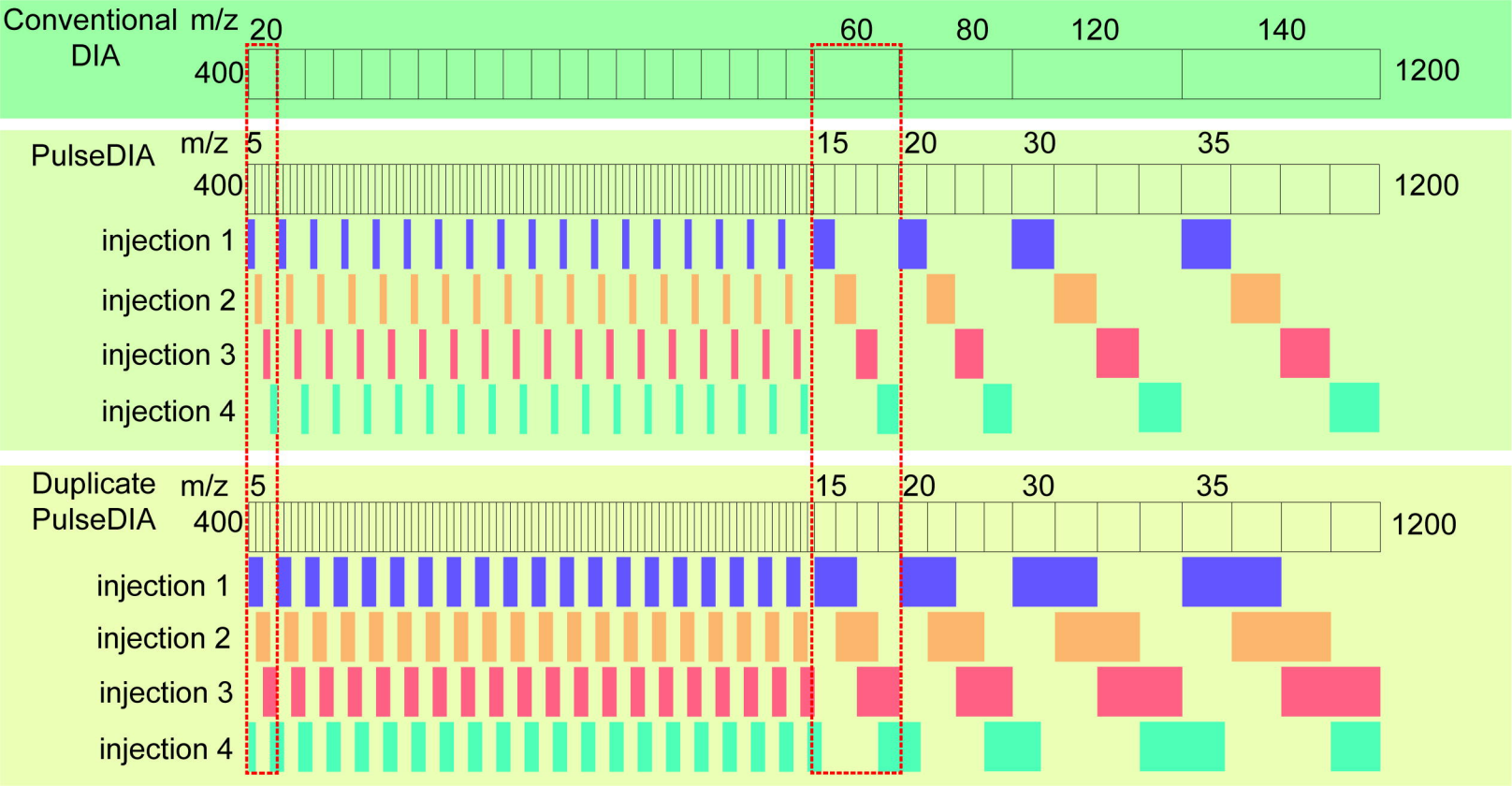
Figure 1. Schematic diagram of PulseDIA with four injections. The MS1 scan range is 400-1200 *m/z*. Conventional DIA-MS has 24 isolation windows. PulseDIA contains 24 isolation windows with 1/4 window width of the conventional one in each pulse injection, and 96 windows in all four injections.

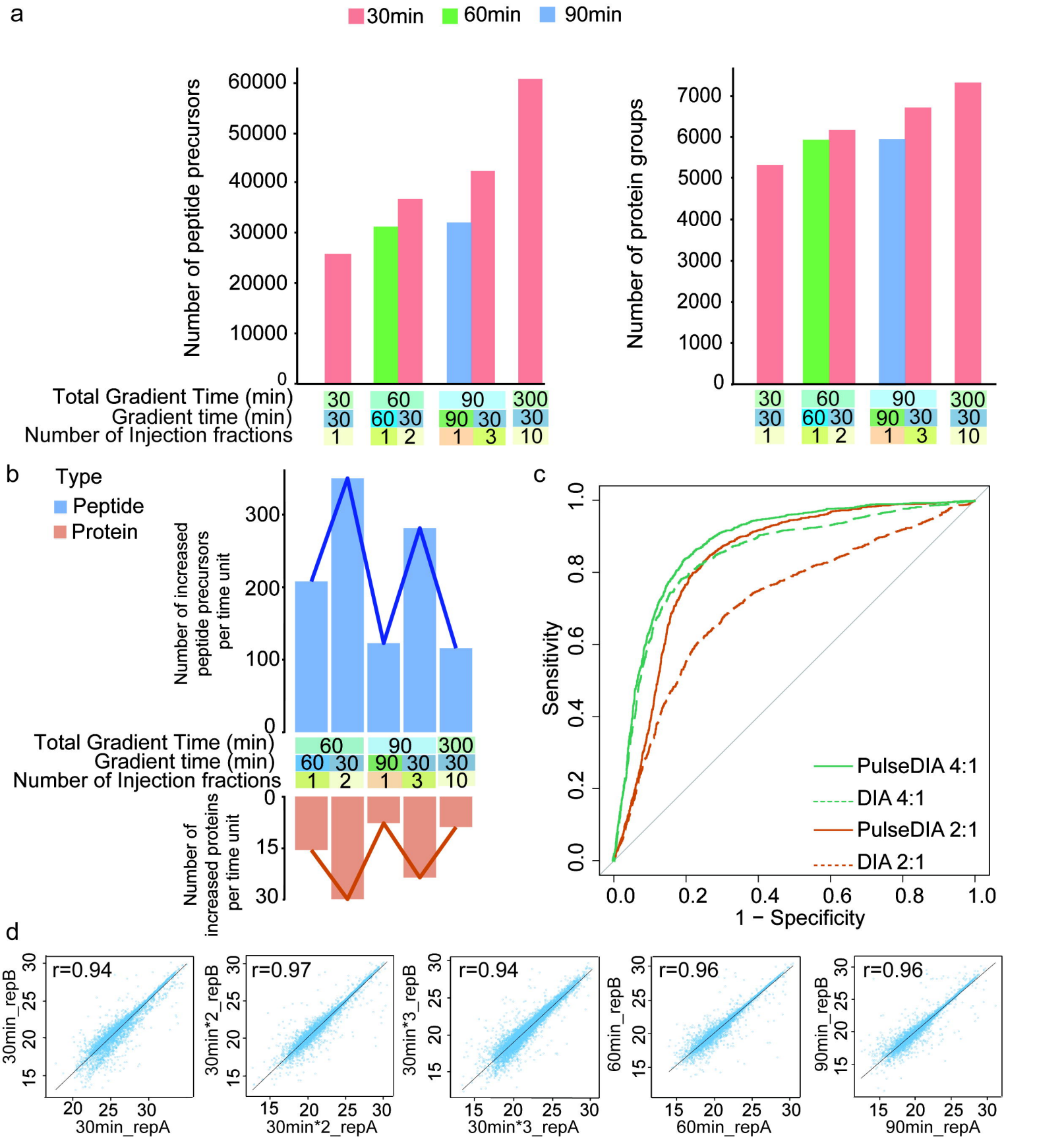
Figure 2. Performance of PulseDIA using HeLa Protein Digest Standard peptides. (a) Number of identified peptide precursors and protein groups using PulseDIA and DIA with different numbers of injections and length of LC gradient. (b) The number of increased peptide precursors and protein groups identified per time unit (min) compared to the conventional DIA of 30 min LC gradient. The blue bar chart represents the peptide precursors, while the orange bar represents protein groups. (c) PulseDIA offers more precise quantification. Sample A (mixture: 1 μg of HeLa digest and 500 ng of *E. coli* digest) was compared to sample B (mixture: 500 ng of HeLa digest and 500 ng of *E. coli* digest) and sample C (mixture: 250 ng of HeLa digest and 500 ng of *E. coli* digest). All samples were analyzed in two technical repeats with a total of 60 min gradient. ROC curves were generated using p values as the classifier, and using the R package called pROC. PulseDIA 2:1 and 4:1 compares sample A:B and A:C, respectively using PulseDIA. DIA 2:1 and 4:1 compares sample A:B and A:C, respectively using DIA. (d) The Pearson correlation between technical replicates for proteins quantified by each method.

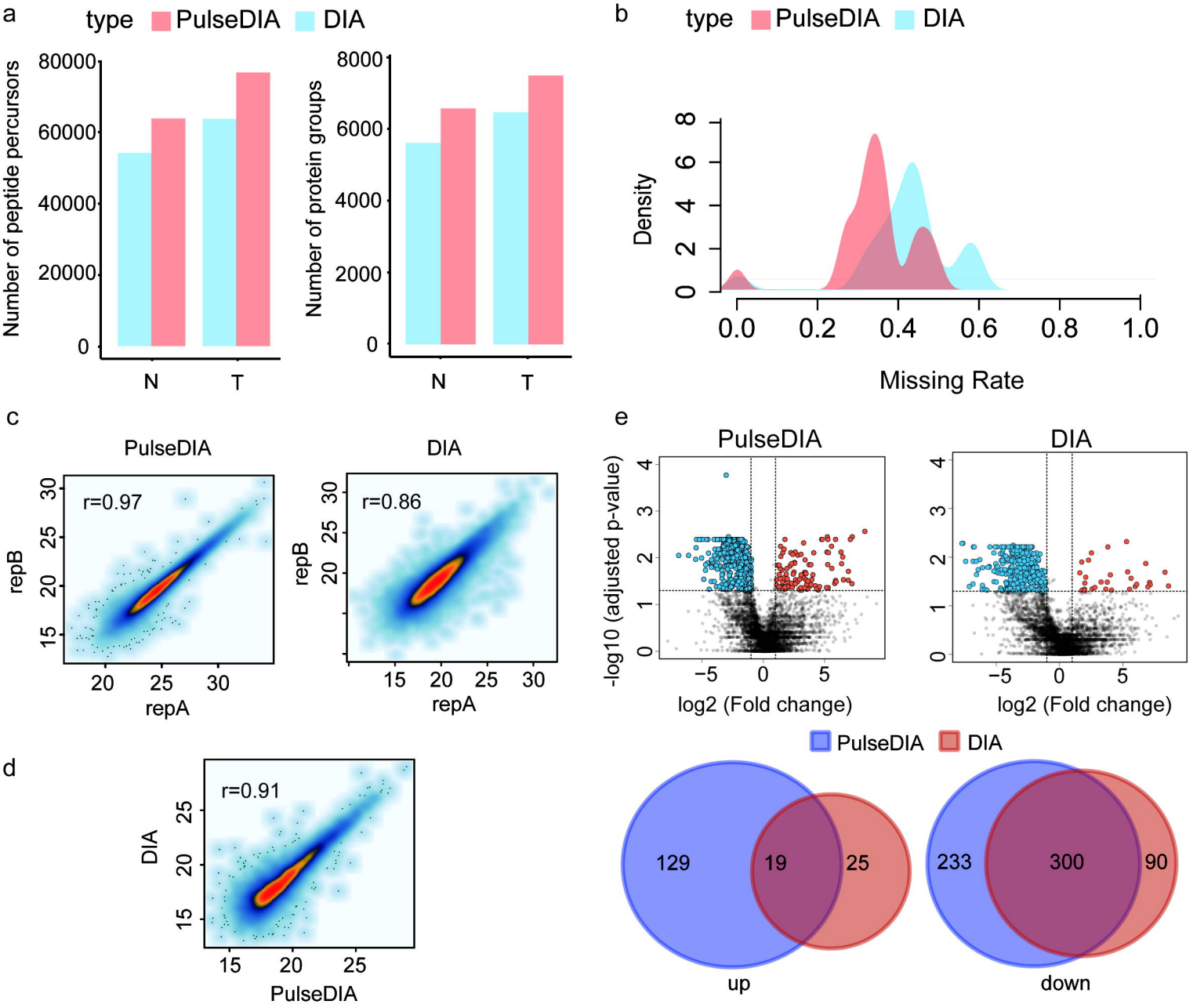
Figure 3. Application of PulseDIA to 18 tissue samples from nine CCA patients. (a) The total number of identified peptide precursors and proteins for all non-tumorous tissues and tumorous tissues by PulseDIA and DIA. N, benign; T, tumor. (b) The missing rate distribution of quantified proteins by two methods. This density curve is drawn from the protein missing rate of each sample in protein matrix. (c) Pearson correlation between technical replicates for proteins using same method. (d) Pearson correlation between quantitative results of the same proteins identified by the PulseDIA and DIA. (e) Volcano plots of regulated proteins for PulseDIA and DIA data sets. Adjusted p values and log2 scaled fold-change values are shown for each protein. Proteins with adjusted p values < 0.05 and $|\log2(FC)| > 1$ are considered as significantly upregulated and downregulated.

Table of Contents graphic




Conventional DIA
m/z
20
60
80
120
140
400
1200
PulseDIA
m/z
5
15
20
30
35
400
1200
injection 1
injection 2
injection 3
injection 4
Duplicate PulseDIA
m/z
5
15
20
30
35
400
1200
injection 1
injection 2
injection 3
injection 4


a
30min
60min
90min
Number of peptide precursors
60000
50000
40000
30000
20000
10000
0
Number of protein groups
7000
6000
5000
4000
3000
2000
1000
0
Total Gradient Time (min)
Gradient time (min)
Number of Injection fractions
30
60
90
300
30
60 30
90 30
30
1
1 2
1 3
10
b
Type
Peptide
Protein
Number of increased peptide precursors per time unit
300
200
100
0
Total Gradient Time (min)
Gradient time (min)
Number of Injection fractions
60
90
300
60 30
90 30
30
1 2
1 3
10
Number of increased proteins per time unit
0
15
30
c
Sensitivity
1 − Specificity
0.0
0.2
0.4
0.6
0.8
1.0
PulseDIA 4:1
DIA 4:1
PulseDIA 2:1
DIA 2:1
d
r=0.94
30min_repB
30min_repA
r=0.97
30min*2_repB
30min*2_repA
r=0.94
30min*3_repB
30min*3_repA
r=0.96
60min_repB
60min_repA
r=0.96
90min_repB
90min_repA
15
20
25
30


a
type
PulseDIA
DIA
Number of peptide percursors
80000
60000
40000
20000
0
N
T
Number of protein groups
8000
6000
4000
2000
0
N
T
b
type
PulseDIA
DIA
Density
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Missing Rate
c
PulseDIA
r=0.97
repB
repA
DIA
r=0.86
repB
repA
d
r=0.91
DIA
PulseDIA
e
PulseDIA
DIA
-log10 (adjusted p-value)
log2 (Fold change)
log2 (Fold change)
PulseDIA
DIA
129
19
25
up
233
300
90
down