Mastering data pre-processing for accurate quantitative molecular profiling with liquid chromatography coupled to mass spectrometry

Mitra, Vikram

Chapter 1. General Introduction

Identification of perturbations in biological systems is the corner stone of biomedical research. Measuring changes at molecular level has allowed biomedical research to understand key physiological and molecular mechanisms in living systems. Molecular profiling studies are useful in discovering biomarkers that can be applied e.g. to monitor disease onset, disease progression, predict treatment efficacy or to identify new proteins that can serve as novel drug targets. To measure perturbations that affect key biological processes at molecular levels, high-throughput molecular profiling platforms are used. Next-generation sequencing and microarrays are used for profiling the genome and transcriptome. DNA and RNA have relatively low chemical diversity with four main nucleotide bases pairs, which can be amplified using polymerase chain reaction. Low complexity and possibility of massive parallel sequencing, allows current genomics technologies to comprehensively profile samples with high sequencing depth and coverage. Whereas, mass spectrometry based molecular profiling coupled with separation techniques such as liquid and gas chromatography, and capillary electrophoresis are used to obtain molecular profile of metabolites, peptides and proteins. These classes of compounds show much higher molecular diversity and complexity compared to DNA and RNA. Although next-generation sequencing allows comprehensive analysis of the genomic and transcriptomics content of cells and tissues, the obtained information does not provide the complete molecular picture of a biological system. Proteins and metabolites are active compounds or results of biological events, and their system wide analysis should be performed and integrated with genomics data to obtain a complete picture of biological states[1,2]. Signals of proteins and metabolites cannot be amplified directly, however, their direct levels in the biological samples can be measured by various profiling technologies. Indirect protein level amplification is possible using aptamers e.g. the Somalogic platform, which allows detection of low abundant proteins in samples with a large concentration range, such as blood[3–5]. The detected indirect signal contains information about concentration levels of proteins but does not provide information about the exact protein sequences or post-translation modifications (PTMs). Key members of the protein-protein interaction networks can be activated or deactivated by PTMs such as glycosylation, phosphorylation, methylation and acetylation[7–9]. Mass spectrometry based proteomics has been fundamental in characterization and quantification of such proteins and PTMs at a system-wide level in various biological systems and conditions[1,2,6–8,10–12].

Complex biological samples containing tens of thousands of proteoforms are successfully analyzed by proteomics approaches using mass spectrometry. Bottom-up proteomics counterintuitively performs enzymatic digestion of proteins into smaller sized peptides, which make the samples even more complex to analyze. However, peptides separated by liquid chromatography (LC), lowers complexity of their detection at specific measurement time. Such complexity reduction is not possible when direct analysis of large intact proteoforms is performed due to lack of suitable separation approach compatible with mass spectrometry. Additionally, large intact proteins cannot be efficiently fragmented, which requires large amount of proteoforms to obtain sufficient ion intensity for large number of fragments. Fragment spectra (or MS/MS spectra) from top-down experiments are too complex for unambiguous identification of the primary amino acid sequence. This has led to the emergence of the bottom-up proteomics technology. In bottom-up proteomics, peptides serve as surrogate compounds for protein characterization and quantitation. Characterization of proteins in mainstream proteomics includes identification of primary amino acid sequence of the resulting peptides.

Data dependent acquisition (DDA) is the most widely used approach for comprehensive profiling of complex proteins and metabolite mixtures in bottom-up experiments[13–15]. DDA LC-MS/MS acquires data in cycles. One cycle starts with acquisition of non-fragmented mass spectrum (MS1 part of LC-MS/MS data) followed by acquisition of fragment MS/MS spectra of the most abundant precursor ions. Depending on the speed of the instrument, 3-20 most abundant ions are fragmented in one duty cycle, and the sampling windows of the non-fragmented ions (precursor ion) is typically 1-2 Da wide. MS1 part of the data is then used for compound quantification in label-free experiments, while the fragment mass spectra is used to ascertain compound identity. The most widely used fragmentation is collision-induced dissociation (CID) and high-energy collision dissociation (HCD) where accelerated intact compound ion beams collide with neutral gases such helium, nitrogen and argon (in vacuum), which produces well defined fragmentation patterns. Electron transfer dissociation (ETD) is another fragmentation approach. In ETD an electron from an aromatic compound acts as a reactive radical anion, which transfers an electron to a positively charged target compound causing its fragmentation (which is different from CID fragmentation). Modern hybrid mass spectrometers can combine CID and ETD fragmentation yielding MS/MS spectra rich in fragments, which may facilitate MS/MS spectra interpretation[16]. A novel acquisition approach called data independent (DIA) LC-MS/MS acquisition is gaining popularity[17–21]. DIA LC-MS/MS acquires data in cycles, however it uses much wider precursor ion isolation window (4-25 Da) and shifts the precursor ion isolation windows

sequentially in order to fragment all precursor ions within a duty-cycle. MS/MS spectra in DDA acquisition mode is used to identify peptide's primary amino acid sequence, while MS1 part of the data is used for accurate peptide and protein quantification.

MS/MS spectra are commonly characterized by database searching. Database search algorithms match m/z list of an experimental MS/MS spectrum to the m/z list of theoretical fragments calculated from peptide sequences similar to m/z of a selected precursor ion. Peptide sequences are generated by performing *in-silico* digestion (using the cleavage rule of the used protease) of the protein sequence database. Following generation of peptide sequences, theoretical lists of fragment ions are generated. The search engine then matches the measured and theoretical list of fragment ions followed by a scores for a match[22,23]. The most common strategy to determine an error in a peptide-spectrum match (PSM) is by performing target-decoy strategy, where false discovery rates for a given score threshold is obtained using the scores of normal and reversed peptide sequences, where the latter are considered as false identifications[24].

Peptide and protein quantitation approaches can be divided into multiple strategies[25,26]. One classification is based on the type of data used for quantification. Spectral counting is simple but less accurate and is based on counting the number of MS/MS spectra that can be attributed to peptides of a given protein[27]. More accurate quantification uses the non-fragmented MS1 information for peptide quantification. Stable-isotope labelling techniques provide peptides, which have the same chemical composition but different stable isotope (^{2}H, ^{13}C, ^{15}N, ^{18}O) constitution. Stable isotope labelling by amino acids in cell culture (SILAC) for example uses metabolic incorporation of amino acids with different isotope composition. Isotope-coded affinity tags (ICAT) and isotope-coded protein labelling (ICPL) are examples of chemical labels which introduces mass differences detectable in MS1. Isobaric tags for relative and absolute quantification (iTRAQ) and tandem mass tags (TMT) employ stable isotope reagent, which reveals the differential labelling upon fragmentation. Stable isotope labelling technology allows peptides to be labelled using different isotopic reagents specific to a sample, hence samples can be mixed. After sample mixing, technical variability of the sample processing and measurement are same for a given compound originating from the different samples, leading to less technical variability and more accurate quantification. However, labelling strategies have caveats such as additional sample processing steps, high cost of labelling reagents and smaller dynamic concentration range due to sample mixing. This resulted in the increase in popularity of the label-free approach in proteomics community, which does not employ any labelling (i.e. label-free quantification). However,

data pre-processing of label-free experiments is complex and challenging as large number of samples are analyzed by independent LC–MS(/MS) runs and information of identical compounds are combined only at the data pre-processing stage at the peak matching step. Matching of analytes across several chromatograms needs to be performed accurately prior to statistical analysis in order to compare the quantities of compounds. Shifts in retention time, m/z an ion intensity of MS1 data need to be corrected prior to any statistical assessment.

The thesis describes novel quality assessment methods to study orthogonality in the retention time domain (separation dimension) and vital pre-analytical factors affecting the ion intensity (readout dimension). Thus following statements form the main goals of the thesis:

- Summary of various data pre-processing steps involved in the treatment of label-free LC-MS(/MS) datasets, obtained for a typical proteomics experiment.
- Describe the MS1 stage of a LC-MS(/MS) dataset as a second order tensor (three dimensional data). Discuss various physio-chemical origins and effects of orthogonality in the two separation dimensions (m/z and retention time) and readout dimension (ion intensity).
- Presentation of a quality assessment approach, which evaluates orthogonality in the retention time dimension for a pair LC-MS(/MS) chromatograms following correction of monotonic shifts.
- Proposition of a method to annotate unmatched spectra based on the concept of “identification transfer” after correction of monotonic shifts between datasets and assess the FDR associated in matching features based on retention time and m/z coordinates between datasets.
- Application of Anova-Simultaneous Component Analysis (ASCA) to determine, which pre-analytical factors influences the ion intensity domain of LC-MS features.

1.1 Data pre-processing steps involved in analysis of label-free LC-MS/MS proteomics datasets

Multiple data pre-processing workflows have been developed to analyse label-free LC-MS(/MS) datasets. Although the inherent algorithmic steps of these routines may vary, the modules of workflows can be clustered into three functional steps, signal processing, quantification and correspondence estimation[10,28–31,27]. Examples of popular open-source workflows are MZmine[32] MaxQuant[33], msInspect[34], VIPER[35], SuperHirn[36], Trans-proteomics-Pipeline (TPP)[37] and OpenMS[27,38,39], while Proteome Discoverer™, PEAKS[40,41] and Progenesis[42] are examples of commercial platforms.

LC-MS(/MS) data contains noise of chemical or electronic origin. This noise is reduced using baseline removal and noise filtering by signal processing algorithms, which improves compound related signal detection and quantification. Examples of noise removal algorithms include morphological filters such as the Top-hat[27] filter which was originally developed to remove noise from binary images. Another example is the asymmetric least squares smoothing algorithm[43] which separates smooth background signal from fast changing compound related signal levels. Numerous other noise-filtering algorithms were also developed, such as moving average, Savitzky–Golay filters[44], entropy-based noise reduction[45] and simple data centroiding, which retains peak local maxima above background signal in each MS1 scan.

Peptides and in general any compound with MS1 signal, can be detected as two dimensional Gaussian or Lorentzian peaks where retention time and m/z values define the coordinates and the ion intensity is the quantitative readout. Quantity of the compound is proportional to the height, area under the curve of a slice or volume of these Gaussian/Lorentzian peaks. One peak in the MS1 stage represents one atomic isotopologue composition of the compounds in a single charge state. One peptide in one charge state is detected as series of isotopologue peaks called features. Isotope pattern of a peptide feature is generally calculated based on the amino acid sequence of the identified peptide or based on “averagine” approach using an averaged atomic composition of peptides[27,46,47]. Peptides are quantified by calculating peak height, area or volume in MS1 data for each isotopologue’s envelope and charge state. Many feature detection algorithms exist, and they mostly fit a 2 or 3 dimensional Gaussian model according to predicted isotopic envelope with direct or iterative approaches[31]. The obtained isotopologue or feature lists are then annotated with peptide identifications and MS(/MS) information[33,38,48].

Signals in MS1 data often have non-linear shifts in both m/z, retention time separation coordinates and ion intensity dimensions. Algorithms for automatic retention time correction are called time alignment algorithms. Mass recalibration algorithms are used for alignment in m/z[31] and normalization procedures are used for "alignment" of ion intensities. From these types of shifts, the most critical is shift in the retention time dimension, since it has a complex physiochemical background and the key causal parameters are difficult to control during data acquisition. Correction of m/z and retention time shifts is required for accurately identifying peaks/features originating from the same compounds in different chromatograms and to correct experimental factors, which influences the ion intensity dimension.

The first-generation of retention time alignment algorithms considered only one separation dimension of the MS1 data and used total ion chromatogram or base peak chromatogram to device the retention time correction function, which was established e.g. using dynamic time warping (DTW)[29,49], parametric time warping (PTW)[50–52] or correlation-optimised time warping (COW)[53,54]. Second-generation time alignment algorithms, such as developed in the Analytical Biochemistry group, included the m/z dimension in the time alignment procedure and therefore used all the two separation dimensions of MS1 data. Besides these algorithms, the implemented noise/signal discrimination such as component detection algorithm (CODA) helped to improve identification of common peaks, which drives the algorithm to device the shift correction function. These developments led to COW-CODA, DTW-CODA and PTW-CODA algorithms[29,54]. Other examples of the second-generation retention time alignment algorithm was developed by Suits *et al*, who used COW and quantitative isotopologue peak lists instead of raw MS data and maximized the sum of overlapping peak volume to device segment-wise the optimal retention time shift correction function[55]. Once correction for shift has been performed in m/z and retention time dimensions, peaks in multiple chromatograms can be clustered using m/z and retention time coordinates with K-means, hierarchical or pose clustering approaches[31,36,39,46].

1.2 Scope of the thesis

Biomarker discovery and statistical analysis of label-free LC-MS(/MS) proteomics profiles requires accurate data pre-processing workflows, to minimise quantification and peak matching errors. In current stage, proteomics is scaling-up for clinical applications where hundreds or even thousands of samples are analysed within a study. For example, the recently launched "Cancer Moonshot" program in US, which is the follow-up of the CPTAC project[56] or the "Proteome of Cancer" or ProCan project in Australia[57]. These projects have the aim to analyse several thousands of tumour samples with the latest LC-MS(/MS)

proteomics profiling platform to complement the large amount of genomics data collected so far to advance cancer research. Data from such types of large sample size studies will be performed in different batches and may be distributed in several laboratories. Experimental design and accurate data processing workflows needs to be developed to analyse and integrate the collected data. In this thesis, I contribute to this aim by developing novel quality assessment approaches for the accurate alignment of MS1 data, required for efficient large-scale data integration and by determining which pre-analytical factors, which effect significantly the LC-MS(/MS) peptide profiles acquired using an experimental design.

The thesis has the following chapters:

Chapter 1 (current chapter) introduces LC-MS(/MS) proteomics profiling technology, describes MS1 data properties and discusses the current LC-MS(/MS) quantitative pre-processing algorithms.

Chapter 2 discusses the physio-chemical origins and effects of non-linear monotonic shifts and orthogonality in the two separation dimensions (m/z and retention time) and quantitative readout ion intensity dimension of MS1 LC-MS(/MS) data. Monotonic shift can be corrected, while the presence of orthogonality can only be assessed, and monotonic shift correction and orthogonality assessment algorithms applied so far in all the three dimensions of MS1 data are discussed in detail in this chapter.

Chapter 3 presents a quality assessment approach, which evaluates the orthogonality between two LC-MS(/MS) chromatograms after correction of monotonic shifts. The chapter presents the detrimental effect of orthogonality on the accuracy of correcting monotonic shifts and the accuracy to predict the retention time of peptide features between chromatograms.

Error rate (false discovery rate, FDR) of peptide-spectrum match, and the lists of identified peptides and proteins in LC-MS(/MS) are strictly controlled by peptide/protein identifications algorithms. **Chapter 4** assesses the FDR of an identification transfer procedure on matching identified and non-identified peptides features between chromatograms using retention time and m/z coordinates in dataset with non-significant orthogonality.

In **chapter 5** we described an approach using Anova-Simultaneous Component Analysis (ASCA) to identify pre-analytical factors that influences LC-MS peptide profiles in human serum samples. In this chapter, we study factors such as, residence time in the autosampler at 4 °C, stopping or not stopping the trypsin digestion with acid, the type

of blood collection tube, different haemolysis levels, differences in clotting times, the number of freeze–thaw cycles, and different trypsin/protein ratios.

In **chapter 6,** I summarize my thesis, and discuss potential follow-up studies on retention time alignment algorithms and on the development of data pre-processing methods aimed to analyse data collected for large-scale proteomics studies.

References

(1) Altelaar, A. F. M.; Munoz, J.; Heck, A. J. R. *Nat. Rev. Genet.* **2012**, *14* (1), 35–48.

(2) Larance, M.; Lamond, A. I. *Nat. Rev. Mol. Cell Biol.* **2015**, *16* (5), 269–280.

(3) Kraemer, S.; Vaught, J. D.; Bock, C.; Gold, L.; Katilius, E.; Keeney, T. R.; Kim, N.; Saccomano, N. A.; Wilcox, S. K.; Zichi, D.; Sanders, G. M.; O'Farrell, P.; Zichi, D.; Eaton, B.; Singer, B.; Gold, L.; Gold, L.; Ayers, D.; Bertino, J.; Bock, A.; Bock, C.; Ellington, D.; Szostak, J.; Tuerk, C.; Gold, L.; Vaught, J.; Bock, C.; Carter, J.; Fitzwater, T.; Otis, M.; Brody, E.; Gold, L.; Famulok, M.; Hartig, J.; Mayer, G.; Gold, L.; Ostroff, R.; Franklin, W.; Gold, L.; Mehan, M.; Miller, Y.; Keeney, T.; Kraemer, S.; Walker, J.; Bock, C.; Vaught, J. *PLoS One* **2011**, *6* (10), e26332.

(4) Gelinas, A. D.; Davies, D. R.; Janjic, N. *Curr. Opin. Struct. Biol.* **2016**, *36*, 122–132.

(5) Gold, L.; Ayers, D.; Bertino, J.; Bock, C.; Bock, A.; Brody, E. N.; Carter, J.; Dalby, A. B.; Eaton, B. E.; Fitzwater, T.; Flather, D.; Forbes, A.; Foreman, T.; Fowler, C.; Gawande, B.; Goss, M.; Gunn, M.; Gupta, S.; Halladay, D.; Heil, J.; Heilig, J.; Hicke, B.; Husar, G.; Janjic, N.; Jarvis, T.; Jennings, S.; Katilius, E.; Keeney, T. R.; Kim, N.; Koch, T. H.; Kraemer, S.; Kroiss, L.; Le, N.; Levine, D.; Lindsey, W.; Lollo, B.; Mayfield, W.; Mehan, M.; Mehler, R.; Nelson, S. K.; Nelson, M.; Nieuwlandt, D.; Nikrad, M.; Ochsner, U.; Ostroff, R. M.; Otis, M.; Parker, T.; Pietrasiewicz, S.; Resnicow, D. I.; Rohloff, J.; Sanders, G.; Sattin, S.; Schneider, D.; Singer, B.; Stanton, M.; Sterkel, A.; Stewart, A.; Stratford, S.; Vaught, J. D.; Vrkljan, M.; Walker, J. J.; Watrobka, M.; Waugh, S.; Weiss, A.; Wilcox, S. K.; Wolfson, A.; Wolk, S. K.; Zhang, C.; Zichi, D.; Zichi, D.; Eaton, B.; Singer, B.; Gold, L.; Pan, S.; Aebersold, R.; Chen, R.; Rush, J.; Goodlett, D.; Service, R.; Liotta, L.; Petricoin, F.; Silberring, J.; Ciborowski, P.; Mitchell, P.; Bell, A.; Deutsch, E.; Au, C.; Kearney, R.; Beavis, R.; Aebersold, R.; Addona, T.; Abbatiello, S.; Schilling, B.; Skates, S.; Mani, D.; Fredriksson, S.; Dixon, W.; Ji, H.; Koong, A.; Mindrinos, M.; Schweitzer, B.; Roberts, S.; Grimwade, B.; Shao, W.; Wang, M.; Borrebaeck, C.; Wingren, C.; Ellington, A.; Szostak, J.; Tuerk, C.; Gold, L.; Gragoudas, E.; Adamis, A.; Jr, E. C.; Feinsod, M.; Guyer, D.; Tarasow, T.; Tarasow, S.; Eaton, B.; Brody, E.; Gold, L.; Famulok, M.; Hartig, J.; Mayer, G.; Gold, L.; Binz, H.; Amstutz, P.; Pluckthun, A.; Eaton, B.; Dewey, T.; Mundt, A.; Crouch, G.; Zyniewski, M.; Eaton, B.; Gugliotti, L.; Feldheim, D.; Eaton, B.; Vaught, J.; Bock, C.; Carter, J.; Fitzwater, T.; Otis, M.; Hopfield, J.; Ninio, J.; Vaught, J.; Dewey, T.; Eaton, B.; Levey, A.; Atkins, R.; Coresh, J.; Cohen, E.; Collins, A.; Chaudhary, K.; Phadke, G.; Nistala, R.; Weidmeyer, C.; McFarlane, S.; Giannelli, S.; Patel, K.; Windham, B.; Pizzarelli, F.; Ferrucci, L.; Nickolas, T.; Barasch, J.; Devarajan, P.; Venturoli, D.; Rippe, B.; Stevens, L.; Coresh, J.; Greene, T.; Levey, A.; Overbeeke, I. van R.; Baan, C.; Hesse, C.; Loonen, E.; Niesters, H.; Pascual, M.; Steiger, G.; Estreicher, J.; Macon, K.; Volanakis, J.; Vanholder, R.; Laecke, S. Van; Glorieux, G.; Doggrell, S.; Ruckman, J.; Green, L.; Beeson, J.; Waugh, S.; Gillette, W.; Jenison, R.; Gill, S.; Pardi, A.; Polisky, B.; Ostroff, R.; Bigbee, W.; Franklin, W.; Gold, L.; Mehan, M.; Schneider, D.; Nieuwlandt, D.; Eaton, B.; Stanton, M.; Gupta, S.; Zichi, D.; Wilcox, S.; Bock, C.;

Schneider, D.; Eaton, B.; Craig, R.; Beavis, R.; Nesvizhskii, A.; Keller, A.; Kolker, E.; Aebersold, R.; Keller, A.; Nesvizhskii, A.; Kolker, E.; Aebersold, R.; Levey, A.; Bosch, J.; Lewis, J.; Greene, T.; Rogers, N. *PLoS One* **2010**, *5* (12), e15004.

(6) Baker, M. *Nature* **2012**, *484* (7393), 271–275.

(7) Zen, Y.; Britton, D.; Mitra, V.; Brand, A.; Jung, S.; Loessner, C.; Ward, M.; Pike, I.; Heaton, N.; Quaglia, A. *EuPA Open Proteomics* **2013**, *1*, 38–47.

(8) Britton, D.; Zen, Y.; Quaglia, A.; Selzer, S.; Mitra, V.; Löbner, C.; Jung, S.; Böhm, G.; Schmid, P.; Prefot, P.; Hoehle, C.; Koncarevic, S.; Gee, J.; Nicholson, R.; Ward, M.; Castellano, L.; Stebbing, J.; Zucht, H. D.; Sarker, D.; Heaton, N.; Pike, I. *PLoS One* **2014**, *9*.

(9) Khoury, G. a.; Baliban, R. C.; Floudas, C. a. *Sci. Rep.* **2011**, *1*, 1–5.

(10) America, A. H. P.; Cordewener, J. H. G. **2008**, 731–749.

(11) Mitra, V.; Smilde, A.; Hoefsloot, H.; Suits, F.; Bischoff, R.; Horvatovich, P. *J. Chromatogr. A* **2014**, *1373*, 61–72.

(12) Ruepp, A.; Waegele, B.; Lechner, M.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; Mewes, H. W. *Nucleic Acids Res.* **2009**, *38* (SUPPL.1), 497–501.

(13) Aebersold, R.; Mann, M. *Nature* **2016**, *537* (7620), 347–355.

(14) Zampieri, M.; Sekar, K.; Zamboni, N.; Sauer, U. *Curr. Opin. Chem. Biol.* **2017**, *36*, 15–23.

(15) Aebersold, R.; Mann, M. *Nature* **2003**, *422* (6928), 198–207.

(16) Shen, Y.; Tolić, N.; Xie, F.; Zhao, R.; Purvine, S. O.; Schepmoes, A. A.; Moore, R. J.; Anderson, G. A.; Smith, R. D. *J. Proteome Res.* **2011**, *10*, 3929–3943.

(17) Scheltema, R. A.; Hauschild, J.-P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. *Mol. Cell. Proteomics* **2014**, *13* (12), 3698–3708.

(18) Huang, Q.; Yang, L.; Luo, J.; Guo, L.; Wang, Z.; Yang, X.; Jin, W.; Fang, Y.; Ye, J.; Shan, B.; Zhang, Y. *Proteomics* **2015**, *15* (7), 1215–1223.

(19) Rosenberger, G.; Koh, C. C.; Guo, T.; Röst, H. L.; Kouvonen, P.; Collins, B. C.; Heusel, M.; Liu, Y.; Caron, E.; Vichalkovski, A.; Faini, M.; Schubert, O. T.; Faridi, P.; Ebhardt, H. A.; Matondo, M.; Lam, H.; Bader, S. L.; Campbell, D. S.; Deutsch, E. W.; Moritz, R. L.; Tate, S.; Aebersold, R. *Sci. data* **2014**, *1* (April 2016), 140031.

(20) 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. *Nat. Protoc.* **2015**, *10* (3), 426–441.

(21) Röst, H. L.; Aebersold, R.; Schubert, O. T. *Automated SWATH Data Analysis Using Targeted Extraction of Ion Chromatograms*; 2016.

(22) Eng, J. K.; Searle, B. C.; Clauser, K. R.; Tabb, D. L. *Mol. Cell. Proteomics* **2011**, *10*, R111.009522-R111.009522.

(23) Hoopmann, M. R.; Moritz, R. L. *Curr. Opin. Biotechnol.* **2013**, *24* (1), 31–38.

(24) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. *Nat. Methods* **2007**, *4*, 923–925.

(25) Zhang, Z.; Wu, S.; Stenoien, D. L.; Paša-Tolić, L. *Annu. Rev. Anal. Chem.* **2014**, *7* (1), 427–454.

(26) Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. *Anal. Bioanal. Chem.* **2007**, *389* (4), 1017–1031.

(27) Nahnsen, S.; Bielow, C.; Reinert, K.; Kohlbacher, O. *Mol. Cell. Proteomics* **2013**, *12* (3), 549–556.

(28) Mortensen, P.; Gouw, J. W.; Olsen, J. V; Ong, S.; Rigbolt, K. T. G.; Bunkenborg, J.; Foster, L. J.; Heck, A. J. R.; Blagoev, B.; Andersen, J. S.; Mann, M. **2010**, 393–403.

(29) Christin, C.; Hoefsloot, H. C. J.; Smilde, A. K.; Suits, F.; Bischoff, R.; Horvatovich, P. L. *J. Proteome Res.* **2010**, *9* (3), 1483–1495.

(30) Griffin, N. M.; Yu, J.; Long, F.; Oh, P.; Shore, S.; Li, Y.; Koziol, J. a; Schnitzer, J. E. *Nat. Biotechnol.* **2010**, *28* (1), 83–89.

(31) Christin, C.; Bischoff, R.; Horvatovich, P. *Talanta* **2011**, *83* (4), 1209–1224.

(32) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinformatics* **2010**, *11*, 395.

(33) Cox, J.; Mann, M. *Nat. Biotechnol.* **2008**, *26* (12), 1367–1372.

(34) Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.; Lin, C.; Chen, J.; Goodlett, D.; Whiteaker, J.; Paulovich, A.; McIntosh, M. *Bioinformatics* **2006**, *22* (15), 1902–1909.

(35) Monroe, M. E.; Tolić, N.; Jaitly, N.; Shaw, J. L.; Adkins, J. N.; Smith, R. D. *Bioinformatics* **2007**, *23* (15), 2021–2023.

(36) Mueller, L. N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M.-Y.; Vitek, O.; Aebersold, R.; Müller, M. *Proteomics* **2007**, *7* (19), 3470–3480.

(37) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. *Proteomics*. 2010, pp 1150–1159.

(38) Weisser, H.; Nahnsen, S.; Grossmann, J.; Nilse, L.; Quandt, A.; Brauer, H.; Sturm, M.; Kenar, E.; Kohlbacher, O.; Aebersold, R.; Malmström, L. *J. Proteome Res.* **2013**, *12* (4), 1628–1644.

(39) Junker, J.; Bielow, C.; Bertsch, A.; Sturm, M.; Reinert, K.; Kohlbacher, O. **2012**.

(40) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337–2342.

(41) Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B. *Molecular & Cellular Proteomics*. 2012, p M111.010587-M111.010587.

(43) Baek, S.-J.; Park, A.; Ahn, Y.-J.; Choo, J.; Tan, H.-W.; Mittermayr, C. R.; Brown, S. D.; Tan, H.-W.; Brown, S. D.; Gemperline, P. J.; Cho, J. H.; Archer, B.; Likar, A.; Vidmar, T.; Hu, Y.; Jiang, T.; Shen, A.; Li, W.; Wang, X.; Hu, J.; Cobas, J. C.; Bernstein, M. A.; Martin-Paster, M.; Tahoces, P. G.; Zhang, Z.-M.; Chen, S.; Liang, Y.-Z.; Baek, S.-J.; Park, A.; Kim, J.; Shen, A.; Hu, J.; Vickers, T.; Wambles, R.; Mann, C.; Mazet, V.; Carteret, C.; Brie, D.; Idier, J.; Humbert, B.; Gan, F.; Ruan, G.; Mo, J.; Baek, S.-J.; Park, A.; Shen, A.; Hu, J.; Zhang, Z.-M.; Chen, S.; Liang, Y.-Z.; Eilers, P. H. C.; Hwang, J.; Choi, N.; Park, A. *Analyst* **2015**, *140* (1), 250–257.

(44) Eilers*, P. H. C. **2003**.

(45) Li, Y.; Qu, H.; Cheng, Y. *Anal. Chim. Acta* **2008**, *612* (1), 19–22.

(46) Kohlbacher, O.; Reinert, K.; Gröpl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M. *Bioinformatics* **2007**, *23* (2), e191-7.

(47) Zhang, J.; Haskins, W. *BMC Genomics* **2010**, *11 Suppl 3* (Suppl 3), S8.

(48) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. *Nat. Methods* **2016**, *13* (9), 741–748.

(49) Kassidas, A.; MacGregor, J. F.; Taylor, P. A. *AIChE J.* **1998**, *44* (4).

(50) Bloemberg, T. G.; Gerretzen, J.; Wouters, H. J. P.; Gloerich, J.; van Dael, M.; Wessels, H. J. C. T.; van

den Heuvel, L. P.; Eilers, P. H. C.; Buydens, L. M. C.; Wehrens, R. *Chemom. Intell. Lab. Syst.* **2010**, *104* (1), 65–74.

(51) Eilers, P. H. C. *Anal. Chem.* **2004**, *76* (2), 404–411.

(52) van Nederkassel, A. M.; Daszykowski, M.; Eilers, P. H. C.; Heyden, Y. Vander. *J. Chromatogr. A* **2006**, *1118* (2), 199–210.

(53) Nielsen, N. P. V; Carstensen, J. M.; Smedsgaard, J. *Journal of Chromatography A*. 1998, pp 17–35.

(54) Christin, C.; Smilde, A. K.; Hoefsloot, H. C. J.; Suits, F.; Bischoff, R.; Horvatovich, P. L. *Anal. Chem.* **2008**, *80* (18), 7012–7021.

(55) Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P. *Anal. Chem.* **2008**, *80*, 3095–3104.

(56) Ellis, M. J.; Gillette, M.; Carr, S. A.; Paulovich, A. G.; Smith, R. D.; Rodland, K. K.; Townsend, R. R.; Kinsinger, C.; Mesri, M.; Rodriguez, H.; Liebler, D. C. *Cancer Discov.* **2013**, *3* (10).

(57) ProCan project http://procan.cancerresearch/About.