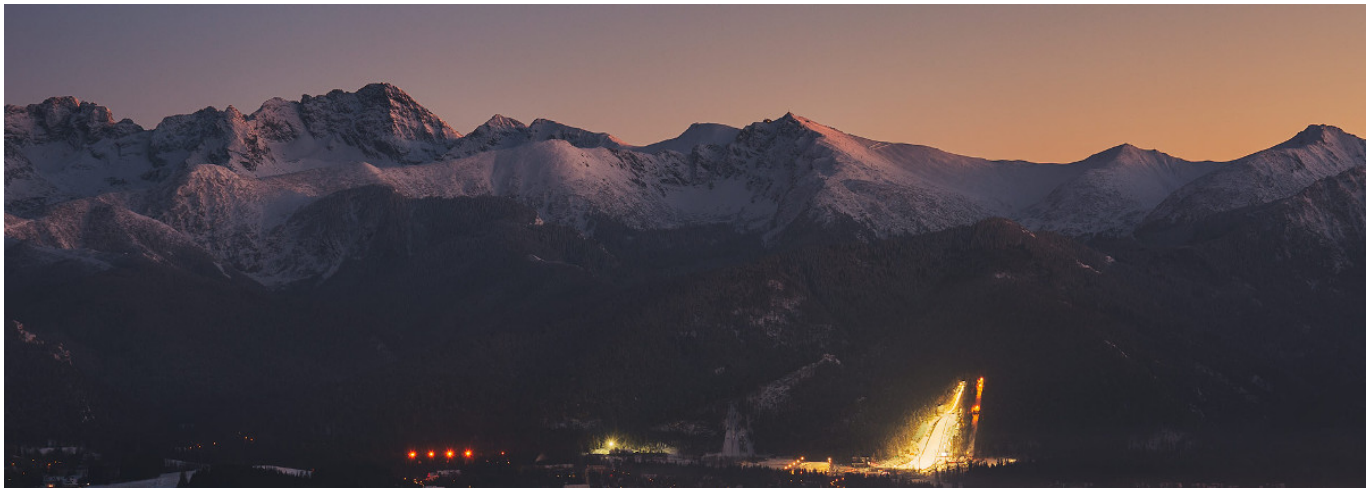

EuBIC Winter School 2019

Programme

15th – 18th January 2019, Hotel Mercure Kasprowy, Zakopane, Poland

https://www.proteomics-academy.org/eubic-winter-school-2019

Contents

Day 1 – 15. January 2019

Educational Day

Time	Title		
09:30 - 10:00	Welcome coffee & Registration		
10:00 - 10:30	Opening Session		
10:30 - 12:00	Introduction to computational mass spectrometry using OpenMS, Part I	Computational introduction into DIA	Label-free quantification: concepts and algorithms
12:00 - 13:00	Lunch Break		
13:00 - 14:30	Introduction to computational mass spectrometry using OpenMS, Part I (continued)	Computational introduction into DIA (continued)	Quantitative proteomics: statistics, clustering and complexes
14:30 - 14:45	Coffee Break		
14:45 - 16:15	Introduction to computational mass spectrometry using OpenMS, Part II	Computational introduction into DIA (continued)	Quantitative proteomics: statistics, clustering and complexes (continued)
16:15 - 16:30	Coffee Break		
16:30 - 18:00	Introduction to computational mass spectrometry using OpenMS, Part II (continued)	Computational introduction into DIA (continued)	Discussion about best practices and common pitfalls

Day 2 – 16. January 2019

DIA and Standards

Time	Title	Speaker(s)
08:35 - 08:40	Morning Welcome and Announcements	
08:40 - 09:25	XCorDIA: a new database search engine to detect geentic variants from DIA data	Brian Searle
09:25 - 10:10	Developing the tools for the personalized medicine revolution: Using mass spectrometry for longitudinal molecular profiling	Hannes Röst
10:10 - 10:30	Coffee Break	
10:30 - 11:15	Label-free Quantification of Complex Proteomes using Ion-Mobility-based DIA	Stefan Tenzer
11:15 - 12:00	Bioinformatics for Proteomics - any open questions?	Martin Eisenacher
12:00 - 13:00	Lunch Break	
13:00 - 13:50	Poster flash talks	
13:50 - 14:20	Sponsor talk	SVA
14:20 - 18:20	Workshop Session (Coffee break at 15:30-15:50)	see workshop descriptions
18:20 - open end	Poster Session & Come Together	

Parallel Workshops

Stefan Tenzer	Quality Control and Benchmarking of Label-Free Quantification Workflows with LFQBench
Martin Eisenacher	The essentials before and after spectrum identification: Selecting the appropriate database and inference strategy
ProFI	Discovering the open-source Proline software suite, a new efficient and user friendly solution for label-free quantification
Thermo	Proteome Discoverer 2.3 Workshop

Day 3 – 17. January 2019

Result Interpretation

Time	Title	Speaker(s)
08:35 - 08:40	Morning Welcome and Announcements	
08:40 - 09:25	STRING – Large-scale integration of data and text	Lars Juhl Jensen
09:25 - 10:10	Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning	Mathias Wilhelm
10:10 - 10:30	Coffee Break	
10:30 - 11:15	Using phosphoproteomics data to study context-specific signalling	Evangelia Petsalaki
11:15 - 12:00	Insights into the multi-functioning proteome	Kathryn Lilley
12:00 - 13:00	Lunch Break	
13:00 - 13:30	Sponsor Talk: Novel DIA Data Analysis Workflow: Integration of De Novo Sequencing and Database Search	BSI
13:30 - 17:30	Workshop Session (Coffee break at 15:30-15:50)	see workshop descriptions
17:30 - 18:00	EuBIC Meeting	All new members are welcome
19:00 - open end	Social Event	

Parallel Workshops

Lars Juhl Jensen	Network visualization with Cytoscape and stringApp
Evangelia Petsalaki (13:30 - 15:30)	SELPHI: using data-driven approaches for analysis of phosphoproteomics datasets
Florian Meier (15:50 - 17:30)	Advanced data acquisition methods with MaxQuant.Live
BSI	PEAKS X: A Complete Solution for Discovery Proteomics with DDA and DIA Support
Matthias Wilhelm	Validation of peptide identifications

Day 4 – 18. January 2019

Innovative methods

Time	Title	Speaker(s)
08:30 - 08:35	Morning Welcome and Announcements	
08:35 - 09:20	Ionbot: a novel, fully data-driven search engine for open modification and mutation searches	Sven Degroeve
09:20 - 09:40	YPIC - Young Proteomics Investigators Club	
09:40 - 10:00	Coffee Break	
10:00 - 10:45	Trapped ion mobility spectrometry: a new dimension for mass spectrometry-based proteomics	Florian Meier
10:45 - 11:00	Announcement: Best Flash Talk, Best Poster Award, Closing Remarks	
11:00	Farewell, lunch boxes and shuttle buses	

Sponsors

Main sponsors:

EuPA
EUROPEAN PROTEOMICS ASSOCIATION

ELSEVIER

Platinum sponsors:

BSi

Gold sponsors:

ProFI
PROTEOMICS

ThermoFisher
SCIENTIFIC
The world leader in serving science

SVA

General sponsor:

Molecular Omics
Editorial Board Chair
Robert Moritz

Associations supporting with grants:

sfeap
société française d'électrophorèse & d'analyse protéomique

BePA

APMA

SEProt
Sociedad Española de Proteómica

Organizers

Dominik Kopczynski Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V. dominik.kopczynski@isas.de	Julian Uszkoreit Ruhr University Bochum julian.uszkoreit@rub.de
Wout Bittremieux University of Antwerp wout.bittremieux@uantwerpen.be	David Bouyssié Institute of Pharmacology and Structural Biology (CNRS) bouyssie@ipbs.fr
Viktoria Dorfer University of Applied Sciences Upper Austria viktoria.dorfer@fh-hagenberg.at	Marie Locard-Paulet Institute of Pharmacology and Structural Biology (CNRS) marie.locard@ipbs.fr
Veit Schwämmle University of Southern Denmark veits@bmb.sdu.dk	Alessio Soggiu Universita Delgi Studi di Milano alessio.soggiu@unimi.it
Sander Willems Ghent University sander.willems@ugent.be	

Talk Abstracts

Day 2 – 16. January 2019

XCorDIA: a new database search engine to detect genetic variants from DIA data

Brian Searle

Single nucleotide polymorphisms and other genomic sequence variants can have profound impact on susceptibility to disease. Even still, most shotgun proteomics workflows focus on detecting canonical protein sequences found in FASTA databases. While proteogenomics methods that combine customized exome sequencing with mass spectrometry are emerging for data dependent acquisition (DDA), data independent acquisition (DIA) approaches frequently rely on curated spectrum libraries that lack sequence variants. Moreover, because most variants result in small retention time and M/Z shifts, these peptides often co-isolate and fragment together in wide DIA precursor isolation windows. Variant peptides produce many of the same fragment ions as canonical peptides and confidently distinguishing different forms is challenging. Moreover, peptide-centric search engines can produce undetectable false positives using these shared ions when searching for low mass PTMs or sequence variants caused by either SNPs, paralogs, or orthologs. We present XCorDIA, a new database search engine that detects and statistically validates PTMS and peptide variants in PEFF databases from DIA data. XCorDIA searches for PTMs and sequence variants by batching peptides that share fragment ions and confirming the presence of specific variants using a PTM/variant detection algorithm similar to PTM site-localization algorithms. We validate XCorDIA using methionine oxidized peptides. Oxidation shifts precursor mass by only +16 Da, which produces mass shifts similar to most SNPs (e.g. +14 Da, V L). Without variant-specific scoring, we find that based on shared fragment ions approximately 1/3rd of oxidized peptides are incorrectly detected at the retention time corresponding to the unmodified form. Finally, we demonstrate how XCorDIA detects sequence variants from ClinVar using clinical amyloidosis samples.

Label-free Quantification of Complex Proteomes using Ion-Mobility-based DIA

Stefan Tenzer

Mass spectrometry-based proteomics greatly benefited from recent improvements in instrument performance and the development of bioinformatics solutions facilitating the high-throughput quantification of proteins in complex biological samples.

Unbiased data-independent acquisition (DIA) strategies have gained increased popularity in the field of quantitative proteomics in the last years, as they provide a complete record of all detectable analytes and facilitates precise label-free quantification of highly complex samples.

The integration of ion mobility separation (IMS) into DIA workflows provides an additional dimension of separation to liquid chromatography-mass spectrometry (LC-MS), and increases the achievable analytical depth of DIA approaches.

From a computational aspect, the increased data complexity provides several opportunities for novel data processing approaches, but is also challenging, as multidimensional solutions for peak picking, alignment and visualization have to be implemented.

Bioinformatics for Proteomics - any open questions?

Martin Eisenacher

Proteomics, especially with mass spectrometry has reached many milestones. Several challenges postulated as being show stoppers have been addressed: identification with limited false positives, quantification, finding "all" gene-coded proteins, modifications (plus localization), usable standard formats. In parallel, instruments and algorithms became more sensitive, more exact and data more sustainable. But there are still some unexplained phenomenons, all-day questions to solve, closed doors to open. For example, the increasing mass accuracy creates new challenges to false-discovery rate estimation. Or, shared peptides could be used for a better quantification. To open the box of pandora - all our method development in mass spectrometry for Proteomics may become obsolete some day.

Day 3 – 17. January 2019

STRING – Large-scale integration of data and text

Lars Juhl Jensen

Methodological advances have in recent years given us unprecedented information on the molecular details of living cells. However, it remains a challenge to collect all the available data on individual genes and to integrate the highly heterogeneous evidence available with what is described in the scientific literature. The STRING database aims to address this by consolidating known and predicted protein–protein association data for a large number of organisms. In my presentation, I will give an overview of the STRING database and describe the general approach we use to unify heterogeneous data, provide comparable quality scores for all evidence types, and automatically mine associations from the biomedical literature.

Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning

Mathias Wilhelm

In mass spectrometry-based proteomics, the identification and quantification of peptides heavily relies on sequence database searching. However, the lack of accurate predictive models for fragment ion intensities impairs the realization of the full potential of this approach. Via ProteomicsDB, reference spectra from the ProteomeTools project and predictions from Prosit are now available, allowing their integration into various proteomic workflows. This talk will showcase some of the many applications in DDA, DIA and PRM workflows.

Using phosphoproteomics data to study context-specific signalling

Evangelia Petsalaki

Phosphoproteomics data provide a snapshot of the phosphorylation-based signaling state of cells. They can therefore be used to dissect the dynamic networks active in a cell in a given condition. Several methods infer context-specific signalling networks from phosphoproteomics data by using as informative priors either existing protein-protein interaction networks or networks from pathway databases. These suffer from severe study bias and therefore data-driven analyses could provide more scope for novel discoveries and an improved understanding of context-specific cell signalling.

To allow non-bioinformaticians to perform data-driven analyses on phospho-proteomics datasets I developed SELPHI, which performs automated data integration and correlation-based network inference. We applied SELPHI to phospho-proteomics data from B cells in variable inhibitor and stimulation conditions, and we identified a novel substrate recognition motif for the Fes kinase. Our follow-up study showed that the motif is recognized by the CSK kinase, and led to explaining the dual oncogenic and tumor suppressive function of Fes.

Currently, we have taken advantage of published phosphoproteomics datasets to generate a data-driven kinase signalling network that can be used as an informative prior for network inference. I will also present preliminary results on a new method that uses phosphoproteomics data to derive context-specific cell signaling networks.

Insights into the multi-functioning proteome

Kathryn Lilley

It is well accepted paradox in biology genome size does not correlate with organismal complexity. In terms of proteins, it could be argued that in higher organisms the proteome is simply too small for the complex functions that it is has to perform [Doolittle, W. F. (2013) Proc. Natl Acad. Sci. USA 110, 5294–5300].

The chemical space that the proteome occupies in higher organisms is vastly expanded by post translational modification, but the numbers and roles of differently functioning proteoforms in a cell are currently uncertain.

In this talk I will review methods that shed light on different functional roles of proteins, from establishing multiple subcellular locations of proteins, to determining additional nucleic acid binding properties of metabolic enzymes. I will also discuss difficulties in trying to determine alternative functional roles for proteins.

Day 4 – 18. January 2019

Ionbot: a novel, fully data-driven search engine for open modification and mutation searches

Sven Degroeve

Modern shotgun proteomics is entirely dependent on accurate search engine tools to match observed spectra to the peptide sequences that generated them. Here we focus on the widely applied approach that is based on a target database that contains all proteins (peptides) expected to be in the sample under consideration. To accommodate for the high computational cost of matching tens of thousands of MS/MS spectra, peptides in the target database are typically considered to be modified by a few of the most common modifications only.

We present Ionbot, a completely new and highly powerful search engine that allows for matching MS/MS spectra against extremely large target databases (allowing thousands of potential protein modifications including mutations). It achieves high processing speeds by implementing a new data-driven approach for selecting candidate peptides for a given MS/MS spectrum. Further, a novel PSM scoring function based on predicted MS/MS spectra is presented as a means to maintain a high degree of sensitivity (at fixed FDR) when handling very large target databases. Ionbot will be demonstrated to perform very well in open modification and mutation searches.

Trapped ion mobility spectrometry: a new dimension for mass spectrometry-based proteomics

Florian Meier

The fast scan speed of time-of-flight analyzers allows adding ion mobility spectrometry as a third dimension of separation. Trapped ion mobility spectrometry (TIMS) is particularly attractive due to its compact design and highly efficient ion utilization. We have recently introduced a novel scan mode termed parallel accumulation – serial fragmentation (PASEF), which synchronizes the release of peptide ions from the TIMS device with the precursor selection in the quadrupole (PMID: 26538118). In data-dependent acquisition, PASEF increases the sequencing by more than 10-fold without any loss in sensitivity (PMID: 30385480). Transferring the PASEF principle to data-independent acquisition could, in principle, capture a much larger proportion of the available ion current as compared with classical DIA, thereby improving sensitivity and acquisition speed several fold. We further demonstrate that peptide collisional cross sections can be readily measured at the scale of 100,000s with high precision. We conclude that TIMS in combination with PASEF is an exciting addition to the technological toolbox in proteomics, with many unique operating modes and applications still left to be explored.

Workshop Abstracts

Day 1 – 15. January 2018

Introduction to computational mass spectrometry using OpenMS

Hannes Röst

We will use the OpenMS library to explore mass spectrometric raw data and data processing basic concepts in mass spectrometry. We will learn about the visualization tools in OpenMS, the scripting capabilities using Python and the internal algorithms and datastructures available. We will also talk about the community and how you can write your own tools in OpenMS and contribute to the project.

Computational introduction into DIA

Maarten Dhaenens and Brian Searle

Label-free quantification: concepts and algorithms

David Bouyssié

Quantitative proteomics, statistics, clustering and complexes

Veit Schwämmle and David Bouyssié

Basic guidelines and methods for visual inspection of quantitative proteomics data, to apply statistical tests, clustering of multivariate data and quantitative assessment of the behavior of protein complexes

Day 2 – 16. January 2019

The essentials before and after spectrum identification: Selecting the appropriate database and inference strategy

Martin Eisenacher und Julian Uszkoreit

In this workshop the participants will learn more about two fundamental pillars of most proteomics studies: the protein databases and the protein inference. We will discuss and show in hands on tutorials which databases are suitable for which analyses and what needs to be considered for the right choice. Furthermore, workflows for performing protein inference using PIA will be explained and teached hands-on.

Attendees should bring their own laptop for the workshop.

Quality Control and Benchmarking of Label-Free Quantification Workflows with LFQBench

Stefan Tenzer

In this workshop, we will introduce the concept of benchmarking label-free quantification workflows using mixed proteome standards.

Applications of LFQBench in instrument benchmarking, quality control and etablishment and validation of quantification workflows will be discussed.

Participants will install and use the open source R-package LFQBench and learn how to interpret the various outputs generated by LFQBench.

Several test datasets will be provided and analyzed.

Please bring your own laptop. Some working experience with R is helpful, but not mandatory for the workshop.

LFQBench can be obtained here: `https://github.com/IFIproteomics/LFQbench`

Proteome Discoverer 2.3 Workshop

Thermo

New Features in PD 2.3

- Statistics and Quantification roll-up strategies for Precursor and Reported based quantification
- Advanced featured and nodes (Cross-linking, Top down)
- Node programming
- Q & A

Discovering the open-source Proline software suite, a new efficient and user friendly solution for label-free quantification

ProFI

DDA Label-free quantification based on precursor ion intensity is a widely used method for quantifying differentially expressed proteins across different conditions or samples. An ideal software solution should allow the production of reliable and comprehensive results, and be flexible enough to allow the integration of existing tools without compromising ease-of-use. To meet these objectives we developed the Proline software, a next-generation tool based on a modular data processing toolbox. This tool constitutes a very interesting alternative to competing solutions, combining robustness, performance, modularity and user-friendliness.

This workshop will be a good opportunity to discover the data processing functionalities of Proline and also the various visualization tools integrated in the Proline-Zero desktop application. After a short introduction of the main software features, we will follow several tutorials aiming at providing a global overview of the tool. During this hands-on session, we will run an the data analysis of a standard dataset composed of an equimolar mixture of 48 human proteins (UPS1, Sigma) spiked at different concentrations into a yeast cell lysate background.

Day 3 – 17. January 2019

Network visualization with Cytoscape and stringApp

Lars Juhl Jensen

The workshop will first provide a quick introduction on the Cytoscape network analysis and visualization tool as well as the Cytoscape stringApp, which makes it easy to import networks from STRING into Cytoscape. Afterwards, we will move on to hands-on exercises, which will teach you how to:

- retrieve networks for proteins or small-molecule compounds of interest
- retrieve networks for a disease or an arbitrary topics in PubMed
- layout and visually style the resulting networks
- import external data and map them onto a network
- perform enrichment analyses and visualize the results
- merge and compare networks
- select proteins by attributes
- identify functional modules through network clustering

If time permits, I will also try to demonstrate how Cytoscape can be used to address some of the challenges that came up during the debate.

SELPHI: using data-driven approaches for analysis of phosphoproteomics datasets

Evangelia Petsalaki

Current phosphoproteomics data analysis pipelines focus mostly on identifying differentially regulated peptides and mapping them on known pathways. This limits our insight around pathways that are well studied and annotated. SELPHI aims to take a data driven approach, to help biologists explore the space less studied in their datasets.

In this workshop I will explain what the aim of SELPHI is and how it works. I will also describe how to generate files for use with SELPHI and will perform a walk through of all the different results that you can acquire using this tool.

Advanced data acquisition methods with MaxQuant.Live

Florian Meier

MaxQuant.Live (www.maxquant.live) is a freely available software framework for real-time monitoring of mass spectrometric data and controlling of the data acquisition. It enables advanced data acquisition strategies on Q Exactive mass spectrometers such as BoxCar (Meier et al., Nat. Methods 2018) and EASI-tag quantification (Virreira Winter et al., Nat. Methods 2018) via a user-friendly graphical interface. Furthermore, it recognizes thousands of peptide precursors in real-time by live re-calibration in three dimensions. In this workshop, you will get familiar with the MaxQuant.Live app store and start generating your own methods, for example a global targeting method for over 20,000 peptides in a single run.

Validation of peptide identifications

Matthias Wilhelm

1. Get predicted spectra from ProteomicsDB
2. Compare results to [proteogenomics/sORF] data using R
3. Investigate effects of pre-processing on spectra similarity

Poster Abstracts

VisioProt-MS: Interactive 2D maps from intact protein mass spectrometry

Locard-Paulet, Marie; Parra, Julien; Albigot, Renaud; Mouton-Barbosa, Emmanuelle; Bardi, Laurent; Burlet-Schiltz, Odile; Marcoux, Julien

Top-down proteomics consists in the analysis of intact proteins using liquid chromatography coupled to mass spectrometry (LC-MSMS). The main advantage of this pipeline over classical bottom-up proteomics is to directly inform on the presence of potential combinations of post-translational modifications, splicing events and/or mutations, thereby providing in-depth characterization of proteoforms. Top-down MS has recently gained momentum, becoming a more high-throughput and quantitative technique. This was allowed by the development of high-resolution mass spectrometers and algorithms allowing signal deconvolution of MS spectra along the chromatographic runs, together with software suites dedicated to proteoform identification.

Visualization of deconvoluted LC-MS data remains a key process in top-down data inspection, and the direct comparison of several LC-MS runs reveals differences in protein footprints between samples/experimental conditions. An increasing number of top-down MS articles can be found in the literature, however only a handful include LC-MS 2D maps.

We thus developed a standalone tool to visualize, inspect and compare the molecular weights (MWs) of eluting proteoforms against their retention times (RT). VisioProt-MS is a user-friendly and highly compatible open source web application that plots and overlays interactive 2D maps from deconvoluted LC-MS run(s). It is designed for dynamic data inspection as well as for creating publication quality figures. VisioProt-MS allows direct input of files from the following bioinformatics tools: RoWinPro, Intact Protein Analysis (BioPharma Finder TM 3.0, Thermo), DataAnalysisTM 4.2 (Bruker), TopFD (TopPIC Suite), ProMex (Informed-Proteomics) for deconvolution of LC-MS data; and Prosight PD (Proteome Discoverer, Thermo), TopPIC, MSPathFinder (Informed-Proteomics) for LC-MSMS data. VisioProt-MS quickly provides an overview of all the detected MWs, reflecting data quality and reproducibility in terms of observed MWs, intensities and RT. It allows comparison of not only multiple LC-MS runs (including from different deconvolution suites), but also LC-MS and LC-MSMS runs of the same sample. Its dynamic features enable to pinpoint potential new proteoforms, quickly reject wrongly assigned Protein Spectral Matches and spot intense proteoforms that remain unassigned. Here we will present the functionalities of VisioProtMS throughout the analysis of different multiproteic complexes and heterogeneous proteins.

Understanding batch-effects through visualisation in proteomics

Willforss, Jakob; Levander, Fredrik

Systematic differences in how samples are handled lead to technical bias in the form of batch effects. Batch effects are common in high-throughput datasets (1), and careful selection of optimal data processing strategies is vital as methods performing poorly for the dataset risks leading to lower sensitivity or even introduce systematic errors causing false positives (2). Data visualisations can reveal unwanted trends in the data, but the degree to which different methods reveal the bias vary between datasets. Investigating the data using multiple types of graphical representations is often needed for a proper understanding of the impact from a batch effect.

This study investigates how well different visualisation methods reveal the presence of batch effects in proteomic datasets. Three types of datasets are studied: Simulated, spike-in and real. The newly developed software NormalyzerDE (3) is used for normalisation and compensation of the batch effect by including a known covariate in the differential expression. Furthermore, we investigate batch-effect compensation methods included in the sva Bioconductor-package including surrogate variable analysis (4) and variants of ComBat (5,6). The batch effects are visualised both on a global level using methods such as principal component analysis (PCA) and p-value histograms and on a feature level investigating protein-specific patterns directly. We study the correspondence between the global- and local- view and relate these to the actual bias.

For most tested datasets, PCA provided a versatile first-view to assess the presence, and relative size of batch effects compared to biological effects but was not always able to pick out batch effects clearly and is unable to assess the impact of batch compensation for methods which do not directly transform the

data. The p-value histogram can be used for assessment as long as a statistical comparison is calculated, but its performance is also dataset dependent. Subsequent visualisations of samples and features proved invaluable for better grasping the potential impact of the batch effect.
We expect that this study will improve the understanding of how batch-effect correction tools perform in proteomic datasets for different types of batch effects and help guide selection of visualisations for understanding the bias. An optimal selection of batch compensation methods would improve post-processing and give a more accurate view of the underlying biology.
(1) Leek, J. et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics
(2) Nygaard, V. et al. (2015) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics
(3) Willforss, J. et al. (2018) NormalyzerDE: Online tool for improved normalization of omics expression data and high-sensitivity differential expression analysis. Journal of Proteome Research
(4) Leek, J. et al. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics
(5) Johnson, WE. et al. (2007) Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics
(6) Zhang, Y. et al. (2018) Alternative empirical Bayes models for adjusting for batch effects in genomic studies. BMC Bioinformatics

MS^2 peak intensity prediction for specific PTMs, fragmentation techniques and instruments

Gabriels, Ralf; Martens, Lennart; Degroeve, Sven

In mass spectrometry-based proteomics, sequence database search engines have proven to be the gold standard in peptide spectrum identification workflows. However, new demands and novel techniques, such as open modification searches and data-independent acquisition, require a higher resolving power to discriminate good from bad search hits. As the traditional search engines do not fully take advantage of the peak intensity information embedded in peptide spectra, doing so can improve the scoring functions. We can obtain peak intensity information for virtually every peptide, by training machine learning algorithms on the vast quantities of data present in public proteomics repositories. The machine learning tool MS^2PIP (MS^2 Peak Intensity Prediction) is already capable of doing so with high accuracy. Nevertheless, many post-translational modifications, fragmentation techniques and instruments influence the peak intensities in such a way, that the general MS^2PIP models underperform when predicting for these special cases.
Because MS^2PIP is a purely data-driven approach, we could train separate models on relevant data sets for phosphorylation, TMT-labeling, TTOF instruments and EThcD fragmentation. With the resulting specific models, we were able to obtain MS^2PIP accuracies as we would expect for normal peptides, even for these special PTMs, fragmentation techniques and instruments.
All MS^2PIP models are available on the user-friendly MS^2PIP web server (https://iomics.ugent.be/ms2pip). Users can upload up to 10 000 peptide sequences simultaneously, for which MS^2PIP predicts MS^2 spectra in just a few seconds. The resulting spectra can be inspected through interactive plots and can be downloaded in both CSV and MGF file formats.

Bioinformatics pipeline for the analysis of proteome data: uncovering surrogate markers of incomplete myocardial reverse remodeling through pericardial fluid proteomics

Trindade, Fábio; Falcão-Pires, Inês; Leite-Moreira, Adelino; Vitorino, Rui

Biomarker discovery has been traditionally pursued by proteomic characterization of easily accessed biofluids such as plasma or urine. The inherent noninvasive nature of collection is irrefutable. Although, there are cases where less easily accessed biofluids can offer a direct window to the diseased organ, serving as a pool to fish biomarkers with higher predicted specificity for a given condition. This is the case of pericardial fluid (PF). If previously considered a mere plasma ultrafiltrate, today's consensus is that PF stores many heart-derived proteins. We hypothesized that screening PF proteome would elicit surrogate prognostic markers for incomplete myocardial reverse remodeling (RR). This is a common phenotype in

aortic valve stenosis (AVS) patients after aortic valve replacement (AVR) surgery and is characterized by limited hypertrophy reversal and/or poor functional recovery. Despite PF's potential as a prognostic platform, an important limitation is the unethical enrollment of healthy control individuals. Therefore, the use of adequate controls must rely on PF from other cardiac pathologies/surgeries lacking the variable/stress of interest. Since we were unable to compare proteome data with healthy individuals, we used coronary artery disease (CAD) patients, without ventricular pressure-overload, as controls. Herein, a specific bioinformatics pipeline to uncover candidate biomarkers for incomplete RR was tailored, which we propose for similar proteomic research.
13 AVS and 6 CAD patients were enrolled. The former patients were divided in complete (n=5) or incomplete RR (n=8) groups, according to the left ventricle mass regression >=4 months after AVR. PF was collected during surgery and its proteome characterized by a shotgun approach using a nanoHPLC-MS/MS system. Data analysis was performed with MaxQuant (version 1.6.1.0) using Andromeda (FDR<1%). Proteins were quantified with the MaxLFQ algorithm (>=2 peptides).
A LFQ intensity matrix was created with normalized proteome data from all subjects (n=19). Data was uploaded to MetaboAnalyst (v4.0), log2-transformed and auto-scaled. A principal component analysis was used to detect outliers (n=2). Differential protein analysis (volcano plots) was performed using a bilateral t-test. Given the lack of healthy controls, we first excluded proteins varying significantly between CAD and AVS patients (1st). This is a key step as CAD is the most prevalent cardiovascular disease, providing us a reasonable level of specificity. We then compared patients with complete and incomplete RR. Proteins quantified in, at least, 5 subjects per group (2nd) and with a fold-change>1.5 (3rd) were selected. Cohen's distance was calculated and proteins consistently up- or downregulated were selected (cohen's distance @95% confidence interval $\neq$0; 4th). Finally, ROC analysis was performed on MetaboAnalyst and proteins with AUC>0.8 were selected (5th). p<0.05 was considered significant.
Overall, 770 proteins were quantified in PF. 20 proteins were found dysregulated between AVS and CAD. 7 proteins were dysregulated during RR. Though, only complement component C8 γ chain, CD5 antigen-like (downregulated) and protein AMBP (upregulated) satisfied cumulatively the proposed five conditions. Therefore, these proteins emerge as candidate markers for incomplete RR. Soon, these results will be validated by immunoblotting and their performance in a multiplex panel will be tested in a larger AVS cohort.

Absolute quantification of influenza A virus proteins using mass spectrometry

Püttker, Sebastian; Behrendt, Ilona; Genzel, Yvonne; Benndorf, Dirk; Reichl, Udo

Even today, there is a high demand for influenza vaccines, seasonal as well as in case of pandemics. Production of influenza virus is either done in chicken eggs or using animal cell culture technology. In the latter, process optimization as well as improvements in virus productivity per cell could boost this technology. Thus, understanding virus replication on the protein level, seeing possible bottlenecks, comparing virus subtypes as well as host cell lines and production methods could facilitate this necessary boost.
The aim of this study is an absolute quantification of viral proteins (HA, NA, M1 and NP) during infection of a MDCK suspension cell line with human influenza A virus (H1N1, A/Puerto Rico/8/34) by mass spectrometry using stable isotope (13C, 15N) labelled signature peptides.
MDCK suspension cells were infected with virus at a multiplicity of infection of 1x10-5 virions/cell and regularly sampled over three days. After centrifugation, cells and supernatant were lysed with SDS buffer, precipitated (acetone) and digested (FASP) using trypsin. Finally stable isotope (13C, 15N) labelled signature peptides were added to the samples. Peptides were submitted to liquid chromatography coupled to a timsTOF Pro mass spectrometer (Bruker Daltonics). Fragment mass data was acquired in data independent mode and processed by the Skyline software.
Initially a selection of tryptic signature peptides was performed based on the protein amino acid sequence of the specific influenza virus strain. Further criteria such as detectability, modifications, miscleavages, and hydrophobicity were used to select the most suitable candidates. This resulted in four peptides for NP, three for NA and HA, and two for M1 protein, respectively. However, using them as isotope labelled surrogates in quantification experiments resulted in different signal abundances, revealing some peptides with low and others with high intensities. The reduction of the ammonium bicarbonate buffer

concentration to 5 mM during the tryptic digestion was found to improve the intensities of the poorly detectable peptides. A further increase in signal intensity was observed when the peptides were added to the complex background of the samples before tryptic digestion, indicating a possible loss of peptides by unspecific binding to the filters. Finally, this improved workflow was used in an initial experiment to measure HA, NA, M1, and NP protein copy numbers during influenza A virus infection of MDCK suspension cells. Results show different dynamics of protein abundances between intra- and extracellular fractions, but also different ratios within the fractions. While in the extracellular fraction the protein ratios roughly corresponded to published data, e.g. M1 as the most abundant protein, protein copy numbers inside the cells indicate a surplus production of NP.
A method forabsolute quantification of influenza virus proteins was established and could help monitoring the dynamics of the viral reproduction cycle on a functional level, which has a high potential to calibrate and improve also existing mathematical models for virus replication on the single cell level. Finally, the new method based on proteome data could help to overcome limitations in process monitoring, i.e. complement assays such as hemagglutination and SRID assay.

Using heavy propionyl to reduce ambiguity in histone annotation

Van Puyvelde, Bart; De Clerck, Laura; Willems, Sander; Daled, Simon; Deforce, Dieter; Dhaenens, Maarten

Histone post-translational modifications (hPTMs) regulate many biological, epigenetic processes. When studied by LCMS, these hPTMs affect both data acquisition and data analysis, therefore dedicated protocols are required. More specifically, histones are often chemically derivatized by e.g. propionylation to block tryptic cleavage of unmodified lysines during sample preparation. However, combining chemical propionyl groups with different biological hPTMs results in many isobaric masses, leading to an increase in ambiguity during data analysis. For histones in particular, multiple variable modifications need to be considered simultaneously during the database search, giving rise to a combinatorial explosion.
Here, we modeled this ambiguity due to propionylation in silico. The results indicate that using heavy propionylation reagents, with the inclusion of e.g. 3 C13 or 5 deuterium, have a more unique mass that potentially reduces ambiguity. However, experimental results are influenced by isotopic impurities of the reagents, impacting both identification and quantification. We conclude that using 3-C13-propionyl minimizes both ambiguity in identification and isotopic skewing of precursors.

DiffPTM: A Shiny/R application to integrate proteomics and PTM-omics data dynamics

Giai Gianetto, Quentin; Chaze, Thibault; Douché, Thibaut; Duchateau, Magalie; Matondo, Mariette

Protein post-translational modifications (PTMs) play a major role in the cellular functions. Changes in PTMs can either cause, or be the result of a disease, making them central to understand the biological functioning of diseases. Label-free quantitation (LFQ) of PTMs by high-resolution mass spectrometry, further analyzed with specific bioinformatics analyses, is a powerful tool to reveal PTM-mediated regulatory networks. Nowadays, robust freeware, such as MaxQuant/Proline/MassChroQ, are available to analyse the large datasets generated by this technique. To find differentially abundant modified peptides between biological conditions, approaches traditionally used in the context of proteins can directly be applied on the measured intensities of modified peptides. However, applying naively such approaches do not answer a crucial question: is the difference of intensities of modified peptides related to the dynamic of their modification, or is it related to the dynamic of the abundance of their belonging protein between compared conditions? This raises statistical issues since we do not observe directly the dynamic of the protein but the one of its peptides in bottom-up proteomics, moreover missing values complicate the problem. Existing software dedicated to statistical analysis of large scale MS-based proteomics experiments either do not propose such analyses (e.g. MSstats, DAPAR/ProStar) or are based on statistical methods that can be questioned (e.g. Perseus). To address this lack of a software dedicated to compare the dynamics of quantified PTMs to their belonging protein, we have developed a new R package linked to a Shiny application called DiffPTM. It proposes functions to statistically compare all the quantified PTMs to its reference proteome between several conditions, and to plot multiple graphs that can directly be included in reports / articles. The Shiny app offers an user friendly interface where unexperimented person (regarding

R programming) can access features of the package just by clicking buttons. Quite importantly, the Shiny app proposes a directed data analysis pipeline to automatically produce PowerPoint reports and Excel files which can be used to provide standardized reports as part of a proteomics platform. These resources are currently used by the Institut Pasteur's proteomics platform and will be soon freely available online for the community.

Global proteome of L3 and L4 Anisakis simplex development stages: TMT-based Quantitative Proteomics. New approach in foodomics.

STRYIŃSKI, ROBERT; MATEOS, JESÚS; BARROS, LORENA; GONZÁLEZ, ÁNGEL F.; PASCUAL, SANTIAGO; GALLARDO, JOSÉ M.; ŁOPIEŃSKA-BIERNAT, ELŻBIETA; MEDINA, ISABEL; CARRERA, MÓNICA

Anisakis simplex is a cosmopolitan parasitic nematode that can cause an illness called anisakiosis. A threat to the health of people all over the world may be the consumption of raw or inadequately prepared fish containing A. simplex larvae, due to their ability to penetrate the mucous membrane of the gastrointestinal tract, as well as severe human allergic reactions. The presence of invasive L3 larvae was documented in 200 species of fish and 25 species of cephalopods around the world, as well as the L4 stage in many species of marine mammals. New culinary trends involving the consumption of raw fish increase the geographical range of parasitic nematodes and the incidence of anisakiosis. Larvae are resistant to freezing, cooking, marinating and salting, which makes them difficult to eliminate.

In this work, using TMT- based (tandem mass tags) quantitative proteomics the global proteome of L3 and L4 development stage of A. simplex was analyzed. The experiment was divided into four stages: (1) extraction of the L3 and L4 larvae proteins, (2) trypsin digestion assisted with high intensity focused ultrasound (HIFU), (3) TMT-isobaric mass tag labeling, and (4) global proteome analysis (LC-MS/MS) of L3 and L4 A. simplex development stages using a LTQ-Orbitrap Elite mass spectrometer.

In this study, we create a reference proteome dataset for each of the two development stages of A. simplex, L3 and L4. Total of 2443 different proteins was identified, where the results showed a high degree of overlap (1542 different proteins) between L3 and L4 of A. simplex. In addition, a high amount of proteins specific only for L3 (330) or L4 (571) were identified and quantified.

Gene ontology (GO) term was performed by the PANTHER classification system to understand the molecular function and biological processes of the identified proteins. Then, KEGG pathway analysis by DAVID 6.8 showed that most of the identified proteins in both stages: L3 and L4 were involved in main metabolic pathways (cel01100), ribosome (cel03010), biosynthesis of antibiotics (cel01130), carbon metabolism (cel01200) and oxidative phosphorylation (cel00190). The most complex nodes of the interaction network in the global proteome were those associated with energy metabolism, regulation of muscle contraction, protein catabolic processes, citrate cycle, aminoacyl-tRNA biosynthesis, and vesicle-mediated transport. The possible interactions were analyzed using the STRING v10.0 software.

Due to the analysis of the specific proteins for L3 and L4 development stages of A. simplex, we identified and characterized many new proteins not yet assigned to this organism. These proteins participate in very important metabolic pathways for parasitic nematodes, which are essential for the development of the parasite. That makes them potential targets in research on antiparasitic substances, as well as may be used for the classification of food and feed contaminants.

This valuable protein repository will add new and significant information to the universal public protein databases and will be very useful for further anisakiosis investigations, and eradication of A. simplex allergens from food, ensuring the safety of the consumers.

Prophane – Metaproteomic Data Analysis and Interpretation Made Simple

Schiebenhöfer, Henning; Schmid, Emanuel; Muth, Thilo; Renard, Bernhard Y.; Riedel, Katharina; Fuchs, Stephan

Metaproteomics or community proteomics is the analysis of proteins in samples composed of multiple organisms. Various issues that already hinder data analysis in proteomics are further complicated in the metaproteomic context. For example, identified peptides can not only stem from homologous proteins in a single organism but also from homologues in different species. As a consequence, metaproteomic

search results largely consist of ambiguous protein identifications that are commonly clustered into protein groups (also called metaproteins). In addition, the interpretation of these search results is complicated because many proteins from non-model species organisms are poorly annotated in reference databases. To simplify the interpretation of protein groups, we developed Prophane, a metaproteomic data analysis software that applies different sequence-based algorithms (DIAMOND BLASTP, HMMER, EMAPPER) to transfer taxonomic and functional annotation from various sources (NCBI NR, SWISSPROT, TREMBL, TIGRFAMs, PFAMs, EGGNOG) to each distinct protein. A lowest common ancestor approach is applied to generate easily interpretable taxonomic metaprotein annotation. Label-free quantitation (NSAF) values are used to visualize sample composition in intuitive and interactive krona plots on both functional and taxonomic level. Prophane is implemented in the pipelining framework Snakemake and, thus, highly scalable and able to process very large data sets in an appropriate period of time. The tool is under active development and will be integrated into an easy-to-use metaproteomic data analysis workflow. This workflow will combine Prophane's features with multiple proteomic search engines, protein identification with flexible grouping and metaproteomics-specific protein quantification.

CUIOS – A tool for visualizing, editing and interpreting single enrichment data

Engler, Alexander; Pielot, Rainer; Höner zu Siederdissen, Christian

In 2016 Kaehne et. al. [1] and later Akondy et. al. [2] used a 2D graph visualization of their meta-analysis data by connecting the meta data tags as vertices via their annotated proteins, weighting these edges according to the number of shared proteins. A force field embedder provided the layout, using the underlying connections. This layouting technique groups similar tags closer to each other than dissimilar ones, helping the researchers recognize co-expressed functions, pathways or processes. In 2009 ClueGo was released as an App for Cytoscape [3], to perform meta analyses and cluster protein lists in a similar fashion [4]. When placing dense clusters or highly similar vertices, two dimensions are limiting the accuracy of the depicted distances, since vertices can't be drawn on top of each other while still providing spatial information. Consequently, the graph will be deformed. Following the ideas of these papers and dealing with the limitations of their solutions, we have developed an easy to use tool for scientists to visualize, interpret and edit their data. CUIOS (Cluster analysis User Interface for all Operating Systems) is using three dimensions to visualize and layout the provided information, overcoming the problems in Kaehne et. al., Akondy et. al. and Bindea et. al.. By clustering and colouring the vertices based on their underlying connections, it further helps researchers to interpret the analysed data. The graph can be rotated, zoomed and edited interactively by using the mouse. When vertices or groups are deleted, the graph automatically updates itself to show how the clusters changed. Nevertheless, CUIOS is written in Java to run on Windows, Linux and MacOS and is optimized for multi CPU systems, using the hardware to its fullest potential while being responsive.

This work was partly funded by the CRC 779 "Neurobiology of Motivated Behavior". References

[1] T. Kaehne, S. Richter, A. Kolodziej, K.-H. Smalla, R. Pielot, A. Engler, F. W. Ohl, D. C. Dieterich, C. Seidenbecher, W. Tischmeyer, M. Naumann and E. D. Gundelfinger, "Proteome rearrangements after auditory learning: high-resolution profiling of synapse-enriched protein fractions from mouse brain,"" Journal of Neurochemistry, vol. 128, no. 1, pp. 124-138, 2016.

[2] R. Akondy, M. Fitch, S. Edupuganti, S. Yang, H. Kissick, K. Li, B. Youngblood, H. Abdelsamed, D. McGuire, K. Cohen, G. Alexe, S. Nagar, M. McCausland, S. Gupta, P. Tata, W. Haining and M. McElrath, "Origin and differentiation of human memory CD8 T cells after vaccination," Nature, no. 552(7685), pp. 362-367, 2017.

[3] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks,"" Genome Res., no. 13(11), pp. 2498-2504, 2003.

[4] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W. Fridman, F. Pagès, Z. Trajanoski and J. Galon, "ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks," Bioinformatics, vol. 25(8), pp. 1091-1093, 2009.

Changes in lamin-associated protein complexes under stress conditions in the Drosophila melanogaster model system.

Pałka, Marta; Tomczak, Aleksandra; Rzepecki, Ryszard

Lamins are nuclear proteins classified as type V intermediate filaments. There are many functions assigned to them so far, including those responsible for maintaining the normal structure of the cell nucleus, regulation of transcription, and the organization of chromatin. Mutations in lamins can cause diseases generally called laminopathies. To date, over 350 mutations in lamin have been identified that results in at least 30 types of diseases. Laminopathies due to very diverse symptoms are an extremely heterogeneous group of diseases. Clinical phenotypes allow laminopathy to be divided into the following groups such as muscular dystrophy, lipodystrophy, neuropathies. In mammalian, we distinguished two types of lamins: A/C- and B-type. A- and C- isoforms can be created from one gene (LMNA) via alternative splicing and the protein product is observed during later stages of embryonic development. B-type lamins are constantly expressed in every cell. There are few isoforms od B-type lamin (the most common are B1 and B2).

To examine the connection between lamins and disorders mentioned above the heat shock induction of stress condition in the Drosophila melanogaster model system was performed. The presence of only two genes coding for lamins (and its high homology to human genes), combined with the simplicity of maintenance and manipulation of Drosophila makes it an excellent candidate for research on laminopathies. The working hypothesis is based on assumption that lamin together with a number of interacting proteins forms complexes, which may change after stress induction. The major aim of this work is to identified potential protein components associated with lamin in normal condition and after heat shock induction and moreover to investigate changes in lamin itself (such as post-translational modifications eg. phosphorylation).

Up to now, preliminary experiments have been carried out to identify protein complexes interacting with lamin. For this purpose, a native immunoprecipitation (IP) of lamin Dm (B-type) was made and the mass spectrometry analysis was performed (LC-MS/MS). Analysis showed that there might be a difference in protein complexes, especially those which functions are connected with protein metabolism, RNA binding or ATP activity. To confirm there results further research is required. Analysis with cross-linked proteins will be performed to determine the exact composition of protein complexes in normal conditions of Drosophila maintenance and after stress condition induction.

VIQoR: an online web service for visually supervised protein inference and protein quantification

Tsiamis, Vasileios; Schwämmle, Veit

Quantitative proteomics measures the changes of protein concentrations between different states of a cell or an organism. Bottom-up proteomics is the most commonly used MS-based approach for protein identification and quantification, according to which proteins are digested into peptides. This process needs to be computationally reverted by inferring proteins from the identified peptides and summarization of peptide abundances for protein quantification. These tasks can be computationally challenging and require powerful methods to avoid too high contribution of wrong and inaccurate peptide abundance measurements to the calculated protein concentrations. We present VIQoR, a user-friendly web service for visually supervised protein inference and protein quantification. The Shiny web interface integrates all the processes involved in protein summarization, along with smart interactive visualization modules to support the common researchers with a straight-forward tool for protein quantification, data browsing and data inspection. We implemented two parsimonious algorithms to solve the protein inference problem, while protein summarization is facilitated by a factor analysis method called fast-FARMS that allows exclusion of missing values and weighted average summarization. The protein inference algorithms create minimal lists of protein groups according to Occam's razor principle with an addition of simple criteria to deal with degenerate peptides. Protein summarization by fast-FARMS assigns individual weights to peptides identifying the same protein group, based on the extracted covariation of their abundances. Low weights correspond to incoherent peptides, which can be eliminated by applying a user-defined weight threshold for the summarization. The tool is implemented in R and its source code will be publicly available soon.

Proteogenomic method applied to RNA-editing investigation

Kuznetsova, Ksenia; Kliuchnikova, Anna; Karpov, Dmitry; Ivanov, Mark; Moshkovskii, Sergei

RNA editing is a posttranscriptional modification done by specific ADAR enzymes. In our work we look for protein products of adenosine-to-inosine substitutions happening in RNA. For these purpose we have adapted the proteogenomic workflow as it is designed for the search of SNP products in proteins.

First, we take the most thorough transcriptomic data and convert it to a proteomic database accounting all known A-to-I substitutions. Inosine in RNA is recognized by the enzymes as guanosine during translation. Then, we map all the resulting peptides to the genomic coordinates using the corresponding resources for a particular organism. After that, we append this database with the "wild" proteins of the same organism taken from UniProt. Finally, this database is used as a fasta file for the proteomic search of the MS/MS spectra.

All the peptide spectra with substitutions pass manual curation. We also use group-specific filtering of the peptides according to target-decoy strategy. The found editing sites undergo validation by other methods such as transcriptomic check of the corresponding sites in RNA and genomic sequencing of the corresponding genes to make sure these substitutions are not encoded in the DNA. The most confident and biologically interesting peptides, then, get validated by Multiple Reaction Monitoring (MRM).

We have successfully applied this method in our recently published work on Drosophila melanogaster [Kuznetsova et al., J. Proteome Res., 2018]. The total number of 68 edited peptides belonging to 59 proteins was identified. Eight of them being shared between the whole insect, head, and brain proteomes. Seven edited sites belonging to synaptic vesicle and membrane trafficking proteins were selected for validation by orthogonal analysis by MRM.

The proteogenomic method allows investigation of RNA-editing sites on proteomic level. It turns out that there are not as many actual edited sites in the proteome as it is in the transctriptome. The reason of such finding partially is that not all the peptides are visible with proteomic methods. Other then that, this difference is caused by the evolution of particular organisms, which is an exciting part of the RNA-editing research and might be developed using multi-omic approach.

Proteogenomics of RNA-editing in model organisms and human

Kliuchnikova, Anna; Kuznetsova, Ksenia; Karpov, Dmitry; Ivanov, Mark; Levitsky, Lev; Moshkovskii, Sergei

Adenosine-to-Inosine (A-to-I) RNA editing is a posttranscriptional modification catalyzed by ADAR enzymes. In most cases, it occurs in nervous tissue, where, as a result of the reaction, adenosine is converted to inosine in particular sites of RNA. We present a proteogenomic study of this phenomenon for three organisms. Proteomic data for Drosophila melanogaster whole body and brain, C57BL/6 mice brain regions and cell cultures, and human brain regions in different stages of development were analyzed for the experiment.

For the fruit fly we have identified 68 edited peptides belonging to 59 proteins. There were two groups of proteins with highly confident interactions. The synaptic signaling group contains proteins which play a role in synaptic transmission, like Syx1A, Syt1, cpx, Syn, AP-2 alpha, endoA, Cadps, and calcium ion channel subunit encoded by CG4587. All proteins from the second group are either components of cytoskeleton or interact with them and take a part in cell transport processes. This group consists of products encoded by zip, alpha-Spec, sls and other.

For the mouse 12 resulting sites were found in 10 proteins encoded by Gria2, Gria3, Gria4, Grm4, neural proteins Flna, Cyfip2 and Cadps. The signal from peptides resulted from A-to-I editing was strongest in early stages of development such as young cell culture of microglia, astrocytes and oligodendrocytes. It was also noted that the neurons such as cortical and cerebellar granule neurons, are subjected to editing more than glial cells such as astrocytes and oligodendrocytes. The resulting sites for human brain data are in good agreement with mice brain editing sites.

All studied organisms had editing sites in proteins connected with synaptic transmission. For mammals the common were modified proteins belonged to AMPA glutamate receptor complex which played a significant role in excitatory synaptic transmission especially in early stages of development. Editing in Cadps required for the Ca2+-regulated exocytosis is common for flies and mice. Identification of A-to-I modifications in these proteins was in good agreement with background works.

Prediction-based reduction of the search space in metaproteomics

Van Den Bossche, Tim

Metaproteomics search databases typically take on enormous sizes since the a priori unknown composition of metaproteomics samples requires the inclusion of proteomes of hundreds to thousands of species that could potentially be found in the samples. A major consequence is that the identification rate in metaproteomics experiments remains drastically below the identification rate in single-species proteomics. Therefore, reducing the database size will not only decrease computation time, but can simultaneously increase identification rate.
To reduce database size, I used predictions from the machine learning algorithm CP-DT. This algorithm, originally intended to predict likely tryptic cleavage sites based on an ensemble of decision trees, has been shown to also be a useful predictor of the likelihood of observing a given peptide in a proteomics experiment. Indeed, if a large database (1.85 million protein sequences) is in silico digested using CP-DT, most peptides are marked as highly unlikely to be observed by the mass spectrometer. Moreover, if the peptide search space is reduced to only the top-35% scoring peptides according to CP-DT, more than 95% of the peptides that were actually observed by the mass spectrometer, are recovered.
From these results I can conclude that the search space can be drastically reduced using CP-DT. Ongoing work will show if this reduction in search space will lead to an increased identification rate, while keeping the false discovery rate (FDR) under control.

The first human protein correlation database uncovers unexpected complexity in protein regulation

Saei, Amir Ata; Zhang, Bo; Beusch, Christian; Sabatier, Pierre; Chernobrovkin, Alexey; Zubarev, Roman A.

Co-expression is routinely used for deciphering gene function through "guilt by association" analysis. We have recently introduced ProTargetMiner, a proteome signature library of 55 anticancer molecules in A549 adenocarcinoma cells encompassing 1,307,859 protein-drug pairs (www.biorxiv.org/content/early/2018/09/18/421115). As the majority of the proteome was perturbed by the compounds, ProTargetMiner provided an opportunity to create the first human protein pairwise correlation database solely based on proteomics data. A 4212 x 4212 matrix was built, from which a high-confidence (FDR<0.001) set of 103,928 positively and 51,137 negatively correlating protein pairs were found representing approximately 1% of the total of 17,740,944 pairs. For every protein pair A-B, we calculated all possible correlations for up- or down-regulation states of A. Five different correlation groups emerged (three positive and two negative), uncovering an unexpected complexity in protein regulation. Most co-regulating proteins (group I) mapped to dense regions of protein interaction networks, such as ribosome and mitochondrial respiratory chain. Besides strong correlations (groups II and III), a surprising number of strong anti-correlations (groups IV and V, 60% of the groups II-III) was found, These findings may contribute to functional annotation of uncharacterized proteins and hint that deeper understanding of cell mechanics is needed for creating a realistic cell model.

Benchmark on recent de novo peptide sequencing tools, including DeepNovo

Altenburg, Tom; Muth, Thilo; Renard, Bernhard Y.

Mass spectrometry is widely used as a high-throughput method in proteomics and metaproteomics studies. Even with this commonality, the results from these studies are often as diverse as the metrics and methods used during down-stream analysis. To overcome this, the community has created gold-standard datasets of MS/MS spectra such as ProteomeTools (Zolg and Wilhelm et al., 2017), a complete set of synthetic human peptides or, in cases of metaproteomics, a 'lab-assembled microbial mixture' (Tanca et al., 2013). Here, we focus on the identification of MS/MS spectra, which involves either a database search or de novo sequencing. This latter method is database independent and allows to identify peptides of unsequenced or even unknown species. Previously, database searches were found superior in terms of accuracy, e.g. assigning the correct peptide sequence. However, recent publications indicate massive improvements regarding de novo sequencing and collectively report higher accuracy and speed. To offer insight on recent implementations of de novo peptide identification, we performed a benchmark on the

most promising, recently developed tools for de novo sequencing; including DeepNovo (Tran et al., 2017). Our results not only show that accuracy indeed made a significant leap for these tools, but also tends to vary between datasets. Furthermore, we found that features, such as the confidence score, assigned by those tools are meaningful and hence constitute promising predictors that may suit as a basis for further down-stream analysis (e.g. protein inference, which relies on the certainty of peptide identifications). With this in mind, we showcast potential future directions for deep learning based approaches relying on a variational autoencoder trained on MS/MS spectra in combination with de novo predictions to further boost sequence identification.

Novel Antidepressant Drug Targets Identification

Yan, Yu; Zhang, Yaoyang; Turck, Christoph W.

Major depressive disorder (MDD) is a common, chronic, recurrent mental illness. However, the pathophysiology and underlying biochemical/molecular events causing MDD remain obscure. Significant drawbacks of currently used antidepressants have prompted the study of their action mechanism and discovery of novel drug targets. Ketamine, a non-competitive NMDA receptor antagonist used for anesthesia, has been found to also have antidepressant activities. A ketamine metabolite, HNK, seems to be critical for its antidepressant effects. However, the mechanism of its antidepressant effect is still unclear. In order to uncover the mechanism of action of antidepressants, we take advantage of mass spectrometry cellular thermal shift assay (MS-CETSA) for the analysis of drug-protein interactions to study the identification of novel protein drug targets. MS-CETSA is combined with TMT-10 plex quantification to get high-throughput proteome data by analyzing thousands of proteins. We found 9 candidates in ketamine treatment gourp and 11 candidates in HNK treatment group, among which pyruvake kinase L/R (PKLR) showed best thermal shift in the assay. It revealed that PKLR could be a potential novel protein target of ketamine or HNK.

MS Annika, a new Search Engine for Identification of Peptides from MS-cleavable Cross-Linkers

Pirklbauer, Georg J.; Stieger, Christian E.; Borgmann, Daniela; Winkler, Stephan M.; Mechtler, Karl; Dorfer, Viktoria

The interest in crosslinking mass spectrometry has risen steadily over the last few years, as has the quality of data and software tools to analyse them [1]. A great improvement came with the development of cross-linkers that are cleavable upon collisional induced dissociation [2]. These linkers enable confident selection of spectra containing cross-linked peptides and provide information for identification.

Here, we present MS Annika, a novel algorithm for the identification of crosslink-spectrum matches (CSMs) from tandem mass spectrometry experiments. MS Annika is specialized on MS cleavable linkers. It is designed to integrate into Proteome Discoverer (Version 2.3), thus eliminating the need for pre-processing steps. The MS Annika algorithm is divided into three stages:

In the first step, MS Annika uses cross-link specific fragment ions, so-called crosslink reporter doublets, to select crosslink spectra. These reporter doublets correspond to the two cross-linked peptides, each of them modified with the heavy and the light part of the cleaved linker. The algorithm also allows for the selection of spectra with incomplete doublets, to increase the number of potential identifications. Based on these doublets, the theoretical precursor masses of the two peptides are identified.

Secondly, a modified version of the MS Amanda [3] database search engine algorithm provides multiple peptide sequences for both precursors. The highest scoring peptides for each precursor are combined to create CSMs. Subsequently, the CSMs are grouped into crosslinks by their cross-linked amino acid site.

The third step comprises a target-decoy based validation. False discovery rates are calculated at CSM as well as crosslink level, resulting in robust identifications.

First results show that MS Annika is able to compete with other tools in the field, both in speed and the number and sensitivity of identifications. For example, we ran both MeroX [4] and MS Annika with default parameters, allowing the DSSO linker to bind to lysine, serine, threonine and tyrosine as well as the protein N-terminus, using carbamidomethylation of C as a static and oxidation of M as a variable modification in a sample with two proteins. From 14708 spectra measured on a Thermo Fischer Q-Exactive

HF mass spectrometer, MeroX identified 234, while MS Annika identified 282 CSMs at an FDR cut-off of 5%.
[1] A. Leitner et al., 'Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines', Trends Biochem. Sci., vol. 41, no. 1, pp. 20–32, Jan. 2016.
[2] A. Sinz, 'Divide and conquer: cleavable cross-linkers to study protein conformation and protein–protein interactions', Anal. Bioanal. Chem., vol. 409, no. 1, pp. 33–44, Jan. 2017.
[3] V. Dorfer et al., 'MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra', J. Proteome Res., vol. 13, no. 8, pp. 3679–3684, Aug. 2014.
[4] M. Götze et al., 'Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis', J. Am. Soc. Mass Spectrom., vol. 26, no. 1, pp. 83–97, Jan. 2015.

MS Ana: A spectral library search engine optimized for high-accuracy fragment ion data

Dorl, Sebastian; Winkler, Stephan; Mechtler, Karl; Dorfer, Viktoria

Spectral library search uses spectrum-to-spectrum matching for the identification of peptides from fragment ion spectra. This approach is now experiencing growing interest in the mass spectrometry community thanks to the increasing number of available spectral libraries. Given a suitable library, using spectrum-to-spectrum matching leads to higher sensitivity and faster processing times than database search1. However, the number of spectral library search engines that are readily available is still small.

We present MS Ana: a spectral library search engine built to take advantage of libraries and experimental data with high-accuracy fragment ions. MS Ana uses an improved scoring function for spectrum-to-spectrum matching in high-accuracy fragment ion data. The scoring uses several different statistical measures that focus on either peak mass or peak intensity and combines all of them to derive a scoring that makes best use of the high-accuracy data. We tested MS Ana performance on a variety of HeLA full cell digest HCD data using the NIST Human HCD spectral library. At FDR 1%, MS Ana identified on average 18.3% more unique peptides than database search with Sequest and 8.8% more unique peptides than to state-of-the-art spectral library search engine SpectraST.

The prominent strategy for controlling FDR in proteomics experiments is the target-decoy approach that carries some issues for spectral library search since decoy library generation is not trivial. MS Ana allows for the generation of new decoy libraries using one of several different algorithms. Decoys can be quickly created for any spectral library independent of library structure or missing fragment annotations.

MS Ana is available as a third-party node for the Thermo Fisher Scientific Proteome Discoverer and can be downloaded free-of-charge from ms.imp.ac.at. Using the Proteome Discoverer software, setting up a search with MS Ana takes only minutes and allows for easy integration with additional analysis tools and existing workflows. References

[1] Zhang, X., Li, Y., Shao, W., & Lam, H. (2011). Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. Proteomics, 11(6), 1075–85.

Proteomic characterization of advanced in-vitro test systems

Salzmann, Eugenia; Bode, Konstantin; Templin, Markus; Poetz, Oliver; Brunner, Thomas; Stoll, Dieter

Following the principles of Replacement, Reduction and Refinement (referred to as the 3R's; (Russell und Burch, 1959)) as the key strategies to provide good ethical, scientific, legal and economic research, in vitro test systems have improved substantially in areas like drug discovery and environmental safety. Although much progress has been made in this field, there still is a need for models in which more levels of organ specific functions may be observed.

Especially three-dimensional systems or co-culture systems, which mimic structural properties observed in vivo, gained increasing interest due to higher predictability, building a long needed bridge between the in vivo situation and traditional in vitro testing. To address this issue we focus on the proteomic characterization of inter alia gastrointestinal organoids as an example for three dimensional cell culture.

By employing a novel immuno-based multiplexed protein profiling technology (DigiWest) detailed information on rare proteomic effects and processes is obtained. Together with a Shotgun-LC-MS approach,

these methods provide a unique concept for studying interactions between all differentiated epithelial cell types of the intestine, and have the potential to provide new insights into cellular signaling processes.

A tool for analysis of kinetics of ETD protein fragmentation

Ciach, Michał Aleksander; Łącki, Mateusz Krzysztof; Miasojedow, Błażej; Lermyte, Frederik; Valkenborg, Dirk; Sobott, Frank; Gambin, Anna

Electron Transfer Dissociation (ETD) is a relatively new protein fragmentation technique, in which the cleavage of a peptide bond is initiated by a rapid neutralization of a positive charge on the protein's backbone. Compared to other standard techniques, like the Collision Induced Dissociation, ETD causes a more uniform fragmentation with fewer losses of amino acid side groups or posttranslational modifications. During the ETD fragmentation process, several other reactions occur, which usually cause charge neutralization without fragmentation. Since such charge loss limits the opportunities for fragmentation, these reactions are usually considered as unwanted. Limiting the impact of side reactions is an art which requires manual fine-tuning of the instrument. Knowledge on the kinetics of the ETD reaction, and how it's influenced by the experimental setting, would allow for faster, easier, and better fragmentation. It would also allow for more accurate computational simulation of mass spectra and identification of unknown proteins. The kinetics of ETD and side reactions has gained some attention, and several models have been proposed. However, they usually require training on massive datasets. In this work, we propose a mathematical model based on Markov jump processes and ordinary differential equations, which does not require data-intensive training procedures. In paticular, it allows to infer reaction rates directly from a single spectrum without any prior training. In our opinion, this is a crucial requirement for a tool aimed at comparing fragmentation in different instrumental settings and of different proteins. Furthermore, the model has only a handful of parameters with a clear interpretation, such as the probability of cleavage at a given residue.
The developed model has been implemented in an open source tool called ETDetective. The presented results have been published in the Journal of Computational Biology.

Characteriziation of proteomic differences in CHO and human multiple myeloma cells

Kretz, Robin; Raab, Nadja; Otte, Kerstin; Fischer, Simon; Poetz, Oliver; Stoll, Dieter; Hauck, Christof

Since their first isolation in 1956, Chinese hamster ovary cells (CHO) were used in various fields, reaching from fundamental research to the production of biologicals in the pharmaceutical industry. In the latter, they are most prominent for their ability to grow in high cell densities and showing glycosylation pattern similar to human cells. Due to the high production cost, extensive downstream processing and the increasing need for biologicals, optimization of the protein secretion of the desired product is continuously conducted.
We will present the first steps of an 'omics approach for the rational cell engineering in CHO cells. The approach is based on comparative studies of the proteome and subproteomes of CHO cells and a human multiple myeloma cell line (JK6L) to reveal differences, both spatial and concentration-dependent, in protein expression which may be correlated to a higher protein secretion in plasma cells. This is done by differential centrifugation of cell lysates followed by LC-MS analysis to identify and quantify the proteins. PCA and machine learning algorithms (SVM, decision trees) are used to map spatial information on the proteins by computing organellar maps from the LC-MS analysis.

Comparison of 5 software packages performing label free quantification: a user's perspective.

Lefeuvre, Bastien; Raffelsberger, Wolfgang; Negroni, Luc

Quantitative proteomics improved by the introduction of the concept of eXtracting Ion Current (XIC) of MS1 peaks in label-free quantitative proteomics. In consequence multiple different algorithms and software-solutions have been developed. In order to find a software development fitting the need of our proteomic platform, we decided to run a benchmark based on two types of samples : a) the commercial human spike-in proteins (UPS) added at 6 different concentrations to a constant base of S. cerevisiae total protein extract and b) a set of samples consisting of H. sapiens, S. cerevisiae and E. coli total protein extracts (HSE) where total protein content as well as H. sapiens proteins were kept constant, while S. cerevisiae and E. coli were added at varied known amounts. This set of samples allowed testing a wide

range of naturally occurring protein abundances and some expected abundance changes ranging from 1.2 to 2.5 fold-change, entering the current limits of detecting differential abundance.
These samples were run as multiple technical replicates on an Orbitrap Elite and were analyzed using Proteome Discoverer v2.2, MaxQuant v1.5.8, MassChroQ v2.2, Proline v1.5 and Progenesis v4.1. Some of these implementations rely on running peptide identification separately while others represent integrated suites, rendering the task of separating the impact of identification and quantitation quite difficult. In order to omit the effect of using different statistical methods by some of the implementations we extracted tables of quantitation at protein level to apply the same filtering, statistical tests and constructed ROC curves in R. The UPS samples (type a) represent an advantage as the number of proteins expected to be identified as variable is precisely known in advance. However, the HSE samples (type b) may be closer to real world settings on our platform where many proteins over a wide range of abundances may vary only to a small degree between given samples. Besides pure comparison of ROC curves for our HSE samples, one should also take the number of quantified proteins in consideration since proteins not detected and/or not quantified by some approaches typically influence apparent precision.

The Wasserstein distance as a dissimilarity measure for mass spectra with application to spectral deconvolution

Skoraczyński, Grzegorz; Ciach, Michał; Miasojedow, Błażej; Majewski, Szymon; Gambin, Anna
We propose a new approach for the comparison of mass spectra using a metric known in the computer science under the name of Earth Mover's Distance and in mathematics as the Wasserstein distance. Under certain assumptions, it can be computed in time linear in the number of distinct peaks in both spectra. We argue that this approach allows for natural and robust solutions to various problems in the analysis of mass spectra. This measure is based on the concept of transporting the ion current between the spectra. The distance between spectra is equal to the total distance in the m/z domain covered by the current. The Wasserstein distance allows to accurately reflect the differences in chemical compositions of the molecules and thanks to computation of transport of ion current allows to match corresponding peaks in the compared spectra, which aids in the detection of differences in elemental composition and chemical structure.
In particular, we show an application to the problem of deconvolution, in which we infer proportions of several overlapping isotopic envelopes of similar compounds. Combined with the previously proposed generator of isotopic envelopes, IsoSpec, our approach works for a wide range of masses and charges in the presence of several types of measurement inaccuracies. To reduce the computational complexity of the solution, we derive an effective implementation of the Interior Point Method as the optimization procedure.
The standard Wasserstein metric allows to deconvolve several highly overlapping isotopic envelopes. However, it requires that all the signal from the experimental spectrum be explained by the theoretical spectra. Due to this requirement, this metric is not robust to chemical noise, and the presence of unexpected molecules can highly perturb the deconvolution results. To account for this, we consider an extension of the metric by allowing to remove unexplained signal from the experimental spectrum, with an appropriate penalty for the amount of signal removed.
Results of deconvolution of both simulated and experimental datasets show that this extension is highly robust to chemical noise, as well as measurement errors in m/z and intensity domain. Notably, the extension does not change the asymptotic time complexity of the algorithm.
The software for mass spectral comparison and deconvolution based on Wasserstein distance is freely available on an open-source licence.

Accelerating fine isotopic structure calculations with IsoSpec 2.0 package

Startek, Michał Piotr; Łącki, Mateusz Krzysztof
IsoSpec is a software library which allows the user to calculate the infinitely-resolved, theoretical spectrum, showing the fine isotopic structure of any molecule whoose chemical formula is known. The capability to quickly calculate a theoretical spectrum of a given molecule is crucial in many mass spectrometry studies, especially in high-resolution mass spectrometry, where the fine isotopic strucutre of the analyte is revealed

by the instrument, and may be used to assist in its identification. Thus, it is useful for identification of the summaric formulas of unknown chemicals, protein sequencing, PTM identification and much more.
The software package is publically available under the open-source BSD licence, it is written in C++, with bindings available for the C, Python and R programming languages. It is also avaialble as a part of the OpenMS mass spectrometry suite.
This poster shortly presents the full capabilities of the software, with particular emphasis on the improvements between the 1.0 and 2.0 version of the software. The algorithmic improvements that led to an order of magnitude speed improvement in the single-threaded algorithm over 1.0 version are presented, as well as the new parallelization schemes and general usability improvements. Last but not least, the poster presents plans and (tenative) timelines for the future development of the software.

Pyteomics 4.0, a proteomics Python library, and what you can do with it

Levitsky, Lev I.; Ivanov, Mark V.; Klein, Joshua; Gorshkov, Mikhail V.
Pyteomics is a freely available open-source Python programming library comprising the building blocks for development of data processing workflows, both for end-user software development and for exploratory data analysis. We announced Pyteomics over five years ago. In this work we summarize its features, including newly added functionality, and provide an overview of data processing tools we have built using Pyteomics.
As a data processing framework, Pyteomics implements reading of proteomics data into Python data types, calculation of peptide and protein properties, and data analysis routines such as linear regression and visualization of fit results.
Specifically, Pyteomics supports reading the following file formats with Python: MGF, mzML, mzXML, ms1; pepXML, MzIdentML, ProteinProphet protXML, X!Tandem XML; featureXML, trafoXML; FASTA, PEFF. A lot of these formats were implemented after the original publication on Pyteomics came out. Additionally, Pyteomics supports writing of MGF and FASTA files, and writing of pepXML and MzIdentML files is implemented with extension packages.
The latest 4.0 release of Pyteomics adds support for indexing to all parsers, which allows fast random access to records rather than sequential iteration only. Additionally, the indexed parsers provide a generic interface for parallelization of user functions applied to records (spectra, PSMs, etc.) in the file.
Apart from data parsing and writing, Pyteomics supports calculation of masses and isotopic composition abundances, retention times, and isoelectric points. It has built-in support for modifications and integrates with the Unimod database. Also, target-decoy approach is supported by providing tools for decoy database generation and decoy-based filtering of search engine results.
Using Pyteomics, we have built a family of open-source data processing tools which we summarize in this work:
IdentiPy - a proteomics search engine with built-in optimization of several key parameters, and complemented with a web-based GUI, IdentiPy Server;
MP score and Scavager, two postsearch validation tools;
AA_stat - a utility for analysis of open search results that helps discover unexpected abundant modifications;
FractionOptimizer - a tool that helps optimize sample fractionation to maximize the proteome coverage;
ms1searchpy - a search engine based on MS1 spectra only;
ms_deisotope - a signal processing, deisotoping, and charge state deconvolution library for reducing mass spectra built on top of Pyteomics;
psims - a mzML and MzIdentML writing library which can produce formatted files from scratch, or pipe transforming functions over Pyteomics readers back to disk.

Retention time alignment driven by partial identifications

Łącki, Mateusz Krzysztof; Diestler, Ute; Tenzer, Stefan
Retention Time Alignment
Modern proteomics offers a variety of methods aimed at characterisation of the molecular composition of bioanalytes.

The ultimate goal of these methods is to achieve highly reproducible measurements.
For this reason, numerous experimental set-ups, such as LC/MS or LC/IMS/MS, have been devised.
The initial step in most of current experiments involves the measurement of retention times of the molecules using liquid chromatography, LC.
In LC, the bioanalyte elutes over time, enhancing detection rates and sensibility of the subsequent methods.
However, the elution profiles can significantly vary even over technical replicates of the same sample.
This limits capabilities of algorithms used for peptide sequencing.
To overcome these issues, several algorithmic approaches have been devised, such as the Match Between Runs algorithm in MaxQuant, or the alignment applied in the IsoQuant software.
Both approaches rely on the combined use of the technical replicates of the experiment.
Signals confidently identified in a series of replicates are used to sequence unidentified signals in the remaining runs.
During this process, the algorithms readjust the retention times, to further help the identification process.
Our current work aims at significantly speeding up the process of alignment through direct use of the information gathered in the sequenced signals.
The approach can be used whenever the space of observed retention times is sufficienly probed by the identified peptides.
With the continuous progress in the methods of analytical chemistry, especially in LC/IMS/MS, the potential limitations of this approach are averted.
The direct use of sequenced signal turns the problem of retention time alignment to a problem akin to nonlinear regression.
We solve the problem by application of robust beta splines.
The proposed algorithm is quick (an alignment of one signal lasts microseconds), thus offering the possibility to perform automated optimization of the method's parameters, through the use of stratified cross-validation.
This liberates the user from specifying the parameters, without compromizing the overall robustness of the method.

Mapping and visualization of the dynamics of histone modifications and their crosstalk

Kirsch, Rebecca; Jensen, Ole Nørregaard; Schwämmle, Veit

Post-translational modifications (PTMs) of histones play a fundamental role in chromatin biology, for instance by regulating gene expression. Chromatin readers, writers and erasers recognize specific combinations of PTMs to regulate chromatin structure and function. This crosstalk between PTMs is not well understood because few experimental platforms can measure multiply modified histones at a large scale. Individual and combinatorial histone modifications can be quantified by middle-down mass spectrometry. To quantify the crosstalk between histone PTMs, an interplay score was developed, which compares the observed co-occurrence of two modifications to the random chance of co-occurrence.
When visualizing histone PTMs and their crosstalk, commonly used hierarchical clustering approaches quickly reach their limits. Since histones generally carry multiple modifications which dynamically change in their abundances, histone PTM datasets consist of multiple layers. This is further complicated when considering experimental designs of different components such as age and tissue. Hence the challenge is to visualize multiple levels of abundance and crosstalk information from a given dataset in a comprehensive way.
We map histone PTM abundances and their crosstalk to coordinates that are invariant to individual histone PTM abundances. This allows to combine data obtained at different time points or from different tissues into one plot while showing the often complex changes in PTM abundances and their crosstalk.
Our visualization framework considerably improves the visualization of the multiple information levels contained in histone PTM datasets. Thus, it simplifies the recognition of complex PTM patterns, helping to disentangle the underlying molecular mechanisms and to identify new features of epigenetic regulation.

Robust summarisation and inference for Label-Free Quantification

Sticker, Adriaan; Goemine, Ludger; Martens, Lennart; Clement, Lieven

Label-Free Quantitative (LFQ) mass spectrometry based workflows have become standard practice in quantitative proteomics for differential expression (DE) analysis of proteins.
Peptides in a mixture are quantified in a first pass through the mass spectrometer (MS1) and selected peptides are identified in a second pass through the mass spectrometer (MS2).
But different peptides from the same protein can have very distinct physio-chemical properties leading to high variability in their MS1 intensities.
Moreover, due to technological constraints not every peptide is selected for identification and low abundant peptides co-eluting with high abundant peptides often get missing.
Peptide-specific effects and context-sensitive missingness makes protein abundance estimation challenging, severely impacting downstream data analysis.
Summarisation methods first aggregate MS1 peptide intensities to protein intensities and DE analysis is done on these protein summaries.
On the other hand, peptide-based models, like MSqRob, allow to test for DE directly from peptide intensities.
By reducing bias and through better uncertainty estimation, they almost always outperform summarisation methods.
However, there are drawbacks toward the use of peptide-based models.
Firstly, fitting peptide based models on complex experimental designs with many samples has an increasingly computational cost.
Secondly, the nonrandom missingness makes it unclear what the correct residual degrees of freedoms are.
Thirdly, MSqRob specifies peptide effects as a random effect in a mixed model, which is often confusing for the non-specialised end-user.
Lastly, MSqRob does not readily provide protein summaries, which are useful for visualisation or downstream processing.
In this work, we use a benchmark spike-in dataset to evaluate recent and often used summarisation strategies.
We discuss why and when these strategies fail compared to the state-of-the-art peptide based model, MSqRob.
We propose a novel summarisation strategy, MSqRobSum, which trains MSqRob in a two-stage procedure circumventing the drawbacks of MSqRob while only suffering a minimal drop in performance.
First, we summarise peptide to protein intensities through robust linear regression, allowing to model peptide-specific effects while still being robust against outliers.
Secondly, the summarised protein intensities are modeled with MSqRob and this is used for DE analysis.
We show huge differences in performance between state-of-the-art summarisation-based strategies depending on the absolute abundance and the fold change of protein expression between conditions, and that MSqRob always outperforms these summarisation strategies.
Our summarisation strategy MSqRobSum, however, has similar performance to MSqRob and only starts to break down at increasingly lower fold changes in protein abundance.
The strategy has several advantages, i.e. summarising peptide to protein intensities reduces the dataset size considerably, speeding up any downstream analysis, determining appropriate degrees of freedom is straightforward, and by specifying our inference model from summarised protein intensities, we avoid the need of peptide random effects, which makes it easier to disseminate our method to a broad audience.
Moreover, MSqRob also has the merit that our analysis framework has become modular.
Indeed, it provides robust protein abundance estimates, which can be used for visualisation and integration in other tools for DE, and MSqRob now also has the functionality to work with summaries from other tools.
This gives our users the additional flexibility to develop modular workflows that are tailored towards their specific applications.

SWATH-MS and pathway analysis show anticancer activity of arachidonic and docosahexaenoic acid monoacylglycerols in colorectal cancer cells

Ortea, Ignacio; González-Fernández, María José; Fabrikov, Dmitri; Ramos-Bueno, Rebeca P.; Guil-Guerrero, José Luis

Background:
Colorectal cancer (CRC) is one of the most common and mortal types of cancer. There is increasing evidence that some polyunsaturated fatty acids (PUFA) are involved in the reduction of cancer risk and progression. Recent studies showed that sn-2 monoacylglycerols (MAGs) exercise specific inhibitory actions on cancer cells through different mechanisms. However, the anticancer effect of PUFA-based MAGs on colorectal cancer has yet to be assessed. Here we investigated the actions of MAGs from two PUFAs, docosahexaenoic acid (DHA) and arachidonic acid (ARA), on CRC human cells, by means of cell assays and SWATH-MS massive quantitative proteomics followed by pathway analysis in order to find out the involved molecular mechanisms.
Methods:
ARA- and DHA-MAG were purified from two commercial oils, DHASCO® (40% DHA) and ARASCO® (40% ARA), using LC. Purified MAGs were added to HT-29 colon cancer cell cultures at several concentrations (50-μ600 M). Cell survival and proliferation, lysis, and apoptosis was assessed by means of MMT, LDH, and caspase-3 assays. Proteome changes produced by each MAG was studied using a SWATH DIA differential proteomics approach, comparing the trypsin-digested proteome of the cells treated with each MAG to control cells. A Triple-TOF 5600+ Q-TOF (Sciex), coupled to nanoLC, was used for all MS analysis. An ad-hoc peptide library was built from the samples using a top 65 DDA LC-MS/MS method followed by peptide and protein identification using ProteinPilot v5.0.1. The SWATH method consisted on the acquisition of the TOF MS/MS of 60 precursor isolation windows of variable width, and data was extracted from the runs using SWATH MicroAPP v.2.0. iPathwayGuide software was used for analyzing the impacted pathways and for Gene Ontology analysis.
Results:
ARA- and DHA-MAG exercised dose- and time-dependent antiproliferative actions. DHA-MAG acted on cancer cells more efficiently than ARA-MAG. 1,882 proteins were quantified in all samples. DHA-MAG produced a deeper effect than ARA-MAG over HT-29 cancer cells proteome (897 vs. 70 differential proteins, p-value$<$0.01 and fold-change$>$2). Pathway analysis revealed that DHA-MAG had a massive effect in the proteasome complex, while ARA-MAG main effect was related to DNA replication.
Conclusion:
Results clearly demonstrated the ability of MAGs to induce cell death in colon cancer cells and suggested a direct relationship between chemical structure and effect. Both MAGs are differentially affecting the whole proteome of HT-29 cells, suggesting that the decrease on cell viability and increase of apoptosis observed should be produced by means of different mechanisms depending on the MAG tested. According to these results, we suggest DHA- and ARA-derived MAGs as candidates that deserve further studies as anticancer effectors for reducing colorectal cancer cell viability.

enhancing matching-between-run with associated uncertainty in retention prediction

Argentini, Andrea; Martens, Lennart

In label-free quantification methods based on MS1 intensities, matching-between-run (mbr) is a technique that address the often-encountered issue of missing data across replicates. The core part of the matching-between-run is the alignment of the retention time (rt) values across the runs in order to be able to predict the RT of a matched peptide in the target run.
In moFF, the mbr is implemented as a combination of pairwise linear models trained using shared peptides across the runs. Although linear models seem to provide an overall good fit of the data, in some cases the elution profile of the peptides is noisier at the beginning and the end of the chromatogram. Moreover, a single value prediction shows also some limitation because a retention time window should be manually associated to the detected peak signal in the raw data.
We have applied a Gaussian Process (GP) model in order to use their probabilistic output to associate an uncertainty window along with the rt value predicted. We have tested our approach on two experiments, where we evaluated the performance in a cross-validation fashion the rt values of identified peptides. Moreover, we have also applied GP with a non-linear kernel to better model the data variation at beginning and the end of the chromatogram with promising result.
In our analysis GP shows comparable performance with the existing linear methods in moFF but with the advantage to have confidence intervals associated to rt predicted reducing the manual imputation of the

retention time window

PLASMA PROTEOME CHANGES IN CHICKEN CHALLENGED WITH LIPOPOLYSACCHARIDE ENDOTOXIN

Horvatić, Anita; Guillemin, Nicolas; Kaab, Haider; McKeegan, Dorothy; O'Reilly, Emily; Bain, Maureen; Kuleš, Josipa; Eckersall, Peter David

The injection of chicken (Gallus gallus domesticus) with bacterial lipopolysaccharide (LPS) is a widely used model for infection and inflammation studies. In order to investigate the immediate innate immune response of chicken to endotoxin-induced inflammation, LPS from Escherichia coli was used to stimulate the response in broiler chickens. The objective of the study was to quantify the changes in chicken plasma proteome after the LPS challenge using tandem mass tag (TMT) label-based high-resolution proteomic analysis. Furthermore, relative protein changes in concentration of established acute phase proteins in chicken plasma, namely serum amyloid A, ovotransferrin and alpha-1-acid-glycoprotein, obtained by proteomic approach were compared to immunoassay-based absolute quantification results.

Plasma from chicken (N = 6) challenged with E. coli (LPS) (2mg/kg body weight) was collected pre (0 h) and at 12, 24, 48, and 72 h post injection along with plasma from a control group (N = 6) challenged with sterile saline. After total protein concentration determination, proteins were reduced, alkylated, acetone-precipitated and labelled with TMT sixplex reagents. Differentially labelled peptides were mixed and analysed using Ultimate 3000 RSLCnano system and Q Exactive Plus mass spectrometer. Identification and relative quantification were performed using Proteome Discoverer. The internal standard (mix of all samples in the study) labelled with TMT m/z 126 was used to compare relative quantification results for each protein between the experiments (sixplexes). Obtained data were analysed using R to determine which proteins were differentially expressed during the different time points (Kruskal-Wallis followed by FDR correction for multiple comparisons). Gene Ontology (GO) terms were analysed by the Cytoscape plugin ClueGO, based on Gallus gallus GO Biological Process database, and refined by REVIGO. As a result, out of 1243 quantifiable peptides identified, 59 related to 19 proteins, including serum amyloid A, ovotransferrin and alpha-1-acid-glycoprotein, showed a significant effect of time post infection in the LPS treated group showing different response patterns. Gene Ontolology terms analyses indicated that pathways related with protein activation cascade (e.g. protein activation cascade, acute-phase response, fibrinolysis, plasminogen activation) and heterotopic cell-cell adhesion were affected by endotoxin challenge.

In conclusion, chicken challenged with bacterial endotoxin demonstrate marked changes to the plasma proteome with both increases and decreases found within 12 hours of challenge. There is potential in this experimental model for biomarker identification, pathophysiological mechanism investigation and as model organism for biomedical research.

Quantitative proteomics reveal novel UBE3A-mediated ubiquitination sites on DDI1 and new insights into its ubiquitin chain type formation

Elu, Nagore; Osinalde, Nerea; Beaskoetxea, Javier; Ramirez, Juanma; Lectez, Benoit; Aloria, Kerman; Rodriguez, Jose Antonio; Arizmendi, Jesus M; Mayor, Ugo

Angelman syndrome (AS) is a rare, complex neurodevelopmental disorder caused by the lack of function in the brain of a single gene, termed UBE3A. This gene encodes an E3 ubiquitin ligase responsible for conferring its substrates with ubiquitin moieties that may greatly influence on the role, regulation and fate of the protein. In AS patients, UBE3A substrates are likely to display a pathologic ubiquitination pattern due to the lack of functional UBE3A in neurons. Therefore, dissection of UBE3A substrate ubiquitination is crucial for a better understanding of the molecular mechanisms underlying this disease. We recently discovered and validated one UBE3A substrate, a proteasomal shuttle protein called DDI1 yet scarcely characterised. In this study, we followed a label free quantitative mass spectrometry-based approach to identify UBE3A-mediated ubiquitination sites and type of ubiquitin linkages on DDI1. We found five novel ubiquitination sites on DDI1, one of them dependent on UBE3A that was further confirmed using site specific mutants. Additionally, we disclosed that UBE3A mainly mediates K48-type ubiquitin linkages on DDI1. Based on the above mentioned results, we propose a mechanism by which K48-type ubiquitination

on DDI1 alters its shuttling function, which ultimately contribute to the proteasomal perturbation that is believed to affect AS patients.

Immunopeptidomics of human tissues using DDA and DIA

Marcu, A.; Bichmann, L.; Backert, L; Kowalewski, D.J.; Freudenmann, L.K.; Kohlbacher, O.; Rammensee, H.G.; Stevanović, S.; Neidert, M.C.

Personalized multi-peptide vaccines are currently discussed intensively for tumor immunotherapy. In order to identify epitopes - short, immunogenic peptides - suitable for eliciting a tumor-specific immune response, human leukocyte antigen (HLA) presented peptides are isolated by immunoaffinity purification from cancer tissue samples and analyzed by liquid chromatography-coupled tandem mass spectrometry (HPLC-MS/MS). To deepen understanding of tissue specific HLA presentation and prevent autoimmunity in the context of epitope-based vaccination, we have assembled a large data set of various healthy human tissues using data dependent (DDA) and independent acquisition (DIA).