Report from Dagstuhl Seminar 19351

Computational Proteomics

Edited by

Nuno Bandeira[1] and Lennart Martens[2]

1 University of California – San Diego, US, bandeira@ucsd.edu
2 Ghent University, BE, lennart.martens@ugent.be

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19351 "Computational Proteomics". The Seminar was originally built around four topics, identification and quantification of DIA data; algorithms for the analysis of protein cross-linking data; creating an online view on complete, browsable proteomes from public data; and detecting interesting biology from proteomics findings. These four topics were led to four correpsonding breakout sessions, which in turn led to five offshoot breakout sessions.

The abstracts presented here first describe the four topic introduction talks, as well as a fifth, cross-cutting topic talk on bringin proteomics data into clinical trials. These talk abstracts are followed by one abstract each per breakout session, documenting that breakout's discussion and outcomes.

An Executive Summary is also provided, which details the overall seminar structure, the relationship between the breakout sessions and topics, and the most important conclusions for the four topic-derived breakouts.

Seminar August 25–30, 2019 – http://www.dagstuhl.de/19351
2012 ACM Subject Classification Applied computing → Bioinformatics
Keywords and phrases computational biology, computational mass spectrometry, proteomics
Digital Object Identifier 10.4230/DagRep.9.8.70

1 Executive Summary

Lennart Martens (Ghent University, BE)
Nuno Bandeira (University of California – San Diego, US)

The Dagstuhl Seminar 19351 'Computational Proteomics' discussed several key challenges of facing the field of computational proteomics. The topics discussed were varied and wide-ranging, and radiated out from the four topics set out at the start.

These four topics were (i) personally identifiable proteomics data; (ii) unique computational challenges in data-independent analysis (DIA) approaches; (iii) computational approaches for cross-linking proteomics; and (iv) the visual design of proteomics data and results, to communicate more clearly to the broad life sciences community. A cross-cutting topic was introduced as well, which focused on proteotyping in clinical trials as it brings many of the previous challenges together, by asking the logical but complex question of how proteomics approaches, data, and associated computational methods and tools can become part of routine clinical trial data acquisition, monitoring and processing.

Computational Proteomics, *Dagstuhl Reports*, Vol. 9, Issue 8, pp. 70–83
Editors: Nuno Bandeira and Lennart Martens
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Based on these initial topics, breakout sessions were organized around proteomics data privacy, dealing with data from DIA approaches, how to best utilize computational approaches to use cross-linking for structural elucidation, and the importance of visualisation of proteomics data and results to engender excitement for the field's capabilities in the life sciences in general. However, these breakout sessions in turn inspired additional breakout sessions on associated topics.

The DIA and cross-linking breakouts both yielded the issue of ambiguity in identification as a cross-cutting topic that merited its own dedicated breakout session. A closley related breakout session, derived from the proteomics privacy and DIA sessions, centered on open modification searches, which are now becoming feasible in proteomics for the first time, but which are also prone to potentially crippling ambiguity issues while raising even more complex privacy issues. The visual design breakout explicitly identified multi-omics data integration as a direct offshoot of its discussions, which led to a dedicated breakout session on this topic as well. Another emerging breakout session concerned public data, which was triggered by both the DIA and cross-linking topics because of their shared need to disseminate their respective specialised data and results in a standardised, uniform, and well-structured manner. Finally, the cross-linking and DIA topics also led to a breakout session on ion mobility, as this technological advance was seen as a key aspect in the future of these technologies.

Each of these breakout sessions had exciting outcomes, and gave rise to future research ideas and collaborations. The proteomics privacy breakout concluded that the field is now ready to delve in more detail into the issues surrounding proteomics data privacy concerns, and that a white paper will be written that can be used to propose policy and to inform the community. The DIA breakout identified three such future tasks: (i) to develop a perspective manuscript that will discuss peptide-centric and spectrum-centric FDR, as well as the effects of shared evidence; (ii) to conduct an experiment for testing DDA versus DIA on the same sample to discover the sampling space for precursors and fragments; and (iii) to conduct a second experiment for understanding target/decoy scoring for different decoy generation models using both synthetic and predicted target/decoy peptides. The cross-linking breakout concluded that a cross-linked ribosomal protein complex should be used as a standardized dataset publicly available to the community, while a 'Minimum Information Requirements About a Cross Linking Experiment (MIRACLE)' was proposed to unify results from many crosslinking tools. The results will also be presented at the Symposium on Structural Proteomics in Göttingen in November 2019. The visual design breakout came up with many fine-grained conclusions, but also with an overall design philosophy which centered on three levels of technical detail, depending on the audience: i) interfaces for deatiled data exploration for experienced consumers; ii) interfaces with minimal technical information, focusing on high-level data for the specific scientific question for novice consumers; and iii) interfaces with only relevant information for clinical decision making (e.g. short list of proteins significantly affected by the disease) for clinicians.

The five offshoot breakouts described above also came to conclusions, and the interested reader is referred to the corresponding abstracts for details.

Overall, the 2019 Dagstuhl Seminar on Computational Proteomics was extremely successful as a catalyst for careful yet original thinking about key challenegs in the field, and as a means to make progress by setting important, high impact goals to work on in close collaboration. Moreover, during the Seminar, several highly interesting topics for a future Dagstuhl Seminar on Computational Proteomics were proposed, showing that this active and inspired community has not yet run out of challenges, nor out of ideas and opportunities!

2 Table of Contents

3 Overview of Talks

3.1 Topic Introduction: Protein Cross-linking

Michael Götze (ETH Zürich, CH), Robert Chalkley (University of California – San Francisco, US), Michael Hoopmann (Institute for Systems Biology – Seattle, US), and Lennart Martens (Ghent University, BE)

The data acquired from cross-linking mass spectrometry (MS) poses specific challenges. These can be split into data processing and analysis concerns on the one hand, and meta-context issues on the other hand. The former revolve around combinatorial problems, due to the large number of possible cross-links that need to be explored.

This in turn leads to ambiguity problems, which are similar to, but exaggerated compared to, classical shotgun proteomics. A further consequence is the apparent limited overlap between different identification algorithms. A last data processing and analysis issue is protein inference, as not only do we need to infer proteins for each linked peptide, we also need to take into account that one of the linked peptides can be very short, which in turn exacerbates the problem.

When it comes to meta-context issues, the first concerns the wide range of the scale: from within-individual protein crosslinking to whole proteome crosslinking. Standard formats are not optimally accommodating right now, and this hampers data dissemination. While initial progress is being made, standard (reference) data sets are not yet sufficiently developed.

Finally, there are opportunities to bring crosslinking results to structural biologists.

3.2 Topic Introduction: Public Proteomics Data: Visual Design and Extraction of Biological Data

Lennart Martens (Ghent University, BE) and Nuno Bandeira (University of California – San Diego, US)

Public proteomics data is currently focused internally primarily. This means that our visualisations are not readily understood outside of our field.

We therefore need a (new) visual design language that can communicate the pertinent information in a readily understood context to specific (outside) users.

We also need to consider the value of public data for novel biological discovery. There is undoubtedly low-hanging fruit there, but we should also look forward at what kind of data (and metadata!) we need to go beyond the low-hanging fruit. In that context, is there something 'special' about proteomics data that makes it more interesting or more relevant to reprocess?

Finally, can we leverage the novel biology that can be found in proteomics data to cement the unique contributions of proteomics data in the context of multi-omics data?

3.3 Topic Introduction: DIA Challenges and Opportunities

Brian Searle (Institute for Systems Biology – Seattle, US) and Maarten Dhaenens (Ghent University, BE)

Data independent acquisition (DIA) mass spectrometry is emerging as a powerful alternative to data dependent acquisition (DDA) and parallel reaction monitoring (PRM). We posit the following questions:

- Are we quantifying peptides at the cost of making detections? How can we convince people to move beyond summing fragments and peptides for protein quant?
- What does FDR mean for DIA? Does target/decoy work the same way as for DDA?
- How to best incorporate ion mobility for DIA? Is establishing peptide overlap for selection potentially more useful than using ion mobility for separation?
- Spectrum-centric versus peptide-centric; and what can we learn from combining these, especially for shared evidence between peptides and PTM positional isomers?
- Can we build and query DIA-based repositories at a raw data level? What types of questions can we answer with "unanticipated" peptide queries across experiments and labs? Is DIA-data (considering it as a digital copy of a sample) even transferable or re-usable between individual labs?
- What can we learn from DIA for proteomics, and how can we apply it to measure metabolites?
- Is it possible to re-use libraries for DIA?

3.4 Topic Introduction: Proteomics Data and Personal Identification

Juan Antonio Vizcaino (EBI – Hinxton, GB)

The detection of genomic variants on a proteome level implies that clinically sensitive proteomics data could be patient-identifiable, and then it should be protected appropriately (for instance, in the context of GDPR guidelines in the European Union).

It is now the right time to assess the state-of-the-art and develop guidelines that are applicable to the community as a whole. Future data management policies for access to human proteomics data in the public domain are part of these efforts.

3.5 Topic Introduction: Proteotyping in Clinical Trials

Bernd Wollscheid (ETH Zürich, CH)

Thinking about and discussing "Clinical Proteotype Analysis" is helpful in order to focus, connect, compare & to make strategic decisions.

- Tumor Profiler project as an example for making such strategic decions
- Which molecular data is 2020 useful in order to support clinical decision-making beyond the current state-of-the-art?
- What is/could be the role of proteotype analysis in the clinical decision-making process?
- In order to particpate in observational & interventional clincial trials we (the proteotype analysis community) needs to make sensible decisions at all levels (sample, sample processing (ID, quant, crosslinking, interactomics, surfaceome analysis etc), data acquisition, data analysis, data visualization, data privacy, data sharing etc
- Matched data generation from the same clinical specimen

4 Working groups

4.1 Working Group Report: Public Proteomics Data

Nuno Bandeira (University of California – San Diego, US), Harald Barsnes (University of Bergen, NO), Frank Conlon (University of North Carolina – Chapel Hill, US), Eric Deutsch (Institute for Systems Biology – Seattle, US), Joshua Elias (Chan Zuckerberg Biohub, US), Rebekah Gundry (University of Nebraska – Omaha, US), Sicheng Hao (Northeastern University – Boston, US), Nils Hoffmann (ISAS – Dortmund, DE), Michelle Kennedy (Princeton University, US), Benoît Kunath (University of Luxembourg, LU), Lennart Martens (Ghent University, BE), Renee Salz (Radboud University Nijmegen, NL), Natalia Sizochenko (Dartmouth College – Hanover, US), Yves Vandenbrouck (CEA – Grenoble, FR), Olga Vitek (Northeastern University – Boston, US), Juan Antonio Vizcaino (EBI – Hinxton, GB), and Bernd Wollscheid (ETH Zürich, CH)

Public availability of proteomics mass spectrometry data has continued to increase to hundreds of terabytes in thousands of datasets from very diverse studies and organisms. However, the lack of metadata describing the samples, experimental design and details of data acquisition and analysis continue to complicate data reutilization and make it difficult for most community members to benefit from the large volume of available data.

This breakout group aimed to compose a vision for the future of public proteomics mass spectrometry data, with a special emphasis on how to make the data most useful to enable clinical and biological discovery.

Two major use cases were proposed to guide the discussion: a) controlled-access clinical proteomics data, typically acquiring larger sample sizes using uniform protocols and featuring extensive sample metadata (often in electronic medical records) and b) open-data research proteomics data, typically acquiring small sample sizes using one or more lab-specific protocols and providing little-to-no metadata describing the study and experiments.

The discussion then focused on incentives that could be implemented to increase the level of annotation of public datasets: i) global associations of expression patterns (e.g., protein expression across tissues, samples-like-mine, etc), ii) offering research tools on the repositories that eventually store the public version of the data (e.g., differential expression, visualization, etc), iii) principal investigator tools (e.g., lab-wide statistics, reports and query features to

support grant writing, etc) and iv) publication guidelines and requirements (e.g., minimal metadata to describe the statistical tests, file and reporting formats, etc).

Finally the group acknowledged the need for example reference datasets illustrating the levels of data and metadata annotation that would be ideal for several classes of technical and biological datasets.

4.2 Working Group Report: Excitement and Visualization

Harald Barsnes (University of Bergen, NO), Michael Götze (ETH Zürich, CH), Rebekah Gundry (University of Nebraska – Omaha, US), Sicheng Hao (Northeastern University – Boston, US), Michael Hoopmann (Institute for Systems Biology – Seattle, US), Michelle Kennedy (Princeton University, US), Benoît Kunath (University of Luxembourg, LU), Lennart Martens (Ghent University, BE), Magnus Palmblad (Leiden University Medical Center, NL), Renee Salz (Radboud University Nijmegen, NL), Natalia Sizochenko (Dartmouth College – Hanover, US), Yves Vandenbrouck (CEA – Grenoble, FR), Olga Vitek (Northeastern University – Boston, US), and Bernd Wollscheid (ETH Zürich, CH)

Creating excitement for proteomics in the community at large, and especially among fellow scientists, is an important goal for the field of proteomics. This starts by figuring out what (mass spectrometry-based) proteomics provides that related technologies do not, and then come up with useful visualizations showing these unique aspects.

Some of the highlighted topics were: i) biological context (e.g. proteins carry out the function, and the majority of drug targets are proteins); ii) the measurement of aggregate events such as post-transcriptional regulation; iii) antibody-independent detection and quantitation of proteins; and iv) the location of post-translational modifications can only be determined by proteomics.

The reasons why proteomics is not well appreciated by other fields was discussed next. This included: i) limited number of known success stories; ii) perceived as inconsistent; iii) higher complexity of the data, i.e. making it harder to interpret; and iv) the high variability in available technologies making it difficult to select the right one.

A couple of solutions were suggested: i) better management of expectations and mindful reporting; ii) create a central source of information on what "proteomics can do for you"; iii) resources promoting proteomics for the non-expert user community; and iv) editorial board members of biology-focused journals should invite contributions focusing on proteomic technologies for non-experts.

Finally, it was suggested that proteomics users can roughly be split into three general categories, all requiring different types of data and visualizations: i) experienced consumers wanting interactive, visual, interfaces for exploring the data in detail; ii) novice consumers requiring minimum amounts of technical information, focusing on what matters to their specific scientific question; iii) clinicians requiring only the information needed to make a clinical decision, e.g. the short list of proteins significantly affected by the disease.

4.3 Working Group Report: Multi-Omics Data Integration (role of proteomics; how to interface)

Pedro Beltrao (EBI – Hinxton, GB), Frank Conlon (University of North Carolina – Chapel Hill, US), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Renee Salz (Radboud University Nijmegen, NL), Brian Searle (Institute for Systems Biology – Seattle, US), Natalia Sizochenko (Dartmouth College – Hanover, US), Yves Vandenbrouck (CEA – Grenoble, FR), Olga Vitek (Northeastern University – Boston, US), Mathias Wilhelm (TU München, DE), Bernd Wollscheid (ETH Zürich, CH), and Roman Zubarev (Karolinska Institute – Stockholm, SE)

Multi-omics data integration can be defined as deriving knowledge from the combination of different Omics measurements that is not possible to obtain from individual data types.

In our discussion, we identified as a major challenge in pursuing such multi-omics studies the increased complexity and skill sets required for the generation and analysis of data of multiple different types. Training is therefore a major issue for developing and carrying out multi-omics studies, and there is a need for combined expertise before a multi-omics project is started and before funding is requested. Most often researchers do not understand what are the opportunities and specific benefits from each Omics technology and how they can be combined in useful ways.

As a concrete step forward, we sought to answer the question, "What can we learn at the interface of omics?" We started to generate a document with pairwise -omics intersections and listed what we thought were a subset of open avenues of research made available by combining different -omics techniques. For some intersections, we added citations to relevant literature that showcases the power of these integrative approaches. This could be developed further into a perspective piece that would help new researchers develop questions to ask about their biological problems and determine which specific methods to focus on learning. This perspective could also serve as an opportunity to generate enthusiasm towards the use of proteomics methods and to highlight where new computational methods are most needed.

4.4 Working Group Report: Open Modification Searches

Robert Chalkley (University of California – San Francisco, US), Nuno Bandeira (University of California – San Diego, US), Lieven Clement (Ghent University, BE), David Creasy (Matrix Science Ltd. – London, GB), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Joshua Elias (Chan Zuckerberg Biohub, US), Michael Götze (ETH Zürich, CH), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), and Juan Antonio Vizcaino (EBI – Hinxton, GB)

There is a long list of tools that have been developed for identifying unanticipated modifications through open mass modification searching. The group discussion mostly focused on three topics: 1. How best to convert an observed modification mass into a named structure; 2. How to assign biological significance to modifications to decide which are worthy of

follow-up; 3. How to create a knowledgebase such that other researchers can learn from previous identifications.

The major outcome from the discussion was a list of recommendations as to how discovered modifications should be reported and stored for community knowledge. This included submitting discovered modifications to Unimod and linking to example spectra in data submitted to a public repository through a universal spectral identifier. A spectral library of these modifications should also be created, although it was acknowledged that there may be challenges in controlling the FDR and FLR in this resource.

4.5 Working Group Report: Ambiguity in Identification (at multiple levels, including FDR)

Lieven Clement (Ghent University, BE), Robert Chalkley (University of California – San Francisco, US), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Joshua Elias (Chan Zuckerberg Biohub, US), and Michael Hoopmann (Institute for Systems Biology – Seattle, US)

Ambiguity is introduced at different levels of the proteomics data analysis workflow. At the level of the identification, protein identification, quantification and differential analysis. Current reporting is driven towards unified results, but ambiguity requires hierarchical classification which is not generally supported by visualization and table schema. Without acknowledging incorrect, though "high-confidence" ambiguous results we risk to draw biological conclusions that may be false.

This breakout group aimed at discussing on a) important types of ambiguity at different levels in the data analysis workflow, b) how these types of ambiguity could be quantified and c) reported. The discussion then focused on challenges in possible strategies and solutions to report more efficiently on ambiguity.

Finally a number of actionable outcomes were selected to be realised on the short term:

1. FDR estimation in identification is currently monopolized by variations on the target decoy approach and it is difficult to publish on alternative ways to estimate the null distribution of false PSMs. With this respect we plan a perspective paper where we will review strategies based on decoys and parametric distributions. We will elaborate on the underlying assumptions of each approach and we will highlight the importance to assess the quality of the approximation of the null distribution within the identification step of the proteomics data analysis workflow.
2. One type of ambiguity that arises in the quantification involves peptides for which the ratios deviate from the proteome ratio. We will develop statistics to prioritise proteins where this type of and we will provide plots to assess the degree of ambiguity.
3. We will assess different types of ambiguity in existing datasets and present the resulting statistics.

4.6 Working Group Report: Cross-linking

Michael Hoopmann (Institute for Systems Biology – Seattle, US), Pedro Beltrao (EBI – Hinxton, GB), Robert Chalkley (University of California – San Francisco, US), David Creasy (Matrix Science Ltd. – London, GB), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Michael Götze (ETH Zürich, CH), Lennart Martens (Ghent University, BE), and Magnus Palmblad (Leiden University Medical Center, NL)

Crosslinking presents many diverse challenges for computational proteomics due to its ever evolving nature of methods and tools. This diversity has hindered development of standardized datasets and workflows.

This breakout session discussed and presented major challenges current to crosslinking data analysis and solutions that will improve upon the field. We focused on three specific tiers for improvement guidelines: the developer, user, and reporting/publication levels.

Within these tiers, primary areas to focus on include improving upon data standardization. Current open standards are poorly implemented, yet there are existing tools such as mzTab that are immediately extensible and will provide greater utility. Additionally, the current paradigm in computational solutions include using tailored datasets. Instead a robust, curated and open dataset utilizing common paradigms and with input from the community was determined to be a better benchmark for algorithm development, and a suggestion was provided.

Additionally, the field suffers greatly from poor validation techniques. For example, current methods applied in standard shotgun proteomics perform poorly when challenged with the sparse datasets in crosslinking. More efforts must be made to explore viable alternatives while expanding the discussion into orthogonal disciplines, such as machine learning.

The discussion concluded by detailing actionable items for which these issues can be addressed. Specifically, a cross-linked ribosomal protein complex could be used as a standardized dataset publicly available to the community, A Minimum Information Requirements About a Cross Linking Experiment (MIRACLE) was proposed extending mzTab that could immediately unify the results from many crosslinking tools.

Furthermore, we expanded the discussion to better define the role of ambiguity in crosslinking, which is essential to understand to improve upon validation.

This topic will be further expanded and discussed in the community at the upcoming Symposium on Structural Proteomics in Göttingen November 3–6, 2019.

4.7 Working Group Report: Ion Mobility

Hannes Röst (University of Toronto, CA), Sebastian Böcker (Universität Jena, DE), David Creasy (Matrix Science Ltd. – London, GB), Eric Deutsch (Institute for Systems Biology – Seattle, US), Maarten Dhaenens (Ghent University, BE), Birgit Schilling (Buck Institute – Novato, US), Brian Searle (Institute for Systems Biology – Seattle, US), Stefan Tenzer (Universität Mainz, DE), Hans Vissers (Waters Corporation – Wilmslow, GB), and Mathias Wilhelm (TU München, DE)

Ion mobility separation (IMS) is an emerging analytical separation technique in various proteomics application areas. It is typically combined with liquid chromatography and mass spectrometry and can provide information on structure, adds an additional dimension of separation, and can have sensitivity benefits.

Discussion topics included IMS principles, hardware configurations, and the computational tools to analyse the multi-dimensional data, which comprises of the following coordinates: retention time, drift time, precursor and product ion m/z, and intensity.

It was concluded that IMS is not widely adopted yet for qualitative and quantitative high-throughput studies. One of the key aspects revolved around the question if IMS provides as solution to the problem of resolving chemeric spectra. It appeared that this is an unresolved question in the field and that further investigation is required. An experiment has been designed to assess the magnitude of this phenomena on two platforms currently available.

A short discussion on data formats showed that current open source data formats are adequate to describe raw data. A document will be distributed describing best practices.

Prediction of CCS is being explored by a number of research groups. However, the benefits of CCS predictions are yet to be determined. Lastly, the impact of IMS on the quantitative performance of label-free quantitation workflows were discussed and experimental designs to evaluate this impact proposed.

IMS has significant potential for multiple applications in MS based proteomics, including structural biology, PTM analysis, and discovery experiments.

4.8 Working Group Report: Data Independent Acquisition

Brian Searle (Institute for Systems Biology – Seattle, US), Sebastian Böcker (Universität Jena, DE), Lieven Clement (Ghent University, BE), Maarten Dhaenens (Ghent University, BE), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Hannes Röst (University of Toronto, CA), Birgit Schilling (Buck Institute – Novato, US), Stefan Tenzer (Universität Mainz, DE), Hans Vissers (Waters Corporation – Wilmslow, GB), and Mathias Wilhelm (TU München, DE)

We discussed several open questions for data independent acquisition (DIA). We first focused on what it means to produce a digital copy of a proteome. DIA produces digital copies of the precursor and fragment ion space, while DDA produces digital copies of only the precursor space. We feel that it might be better to talk about DIA as creating a "deterministic copy" or

"consistent copy" rather than a "digital copy" to emphasize that it is not a "comprehensive copy" and it doesn't contain every possible peptide.

We have an interest in determining some statistics about DIA to determine the parameters in which it works better than DDA. In particular, we are interested in asking:

- How many of the precursors do we see in MS1, how many of them trigger MS2 spectra?
- Are we still under-sampling the precursor space?
- Are the MS2 spectra of poor quality due to triggering early and getting poor quality spectra?

We plan to conduct an experiment to probe these questions.

We then discussed the effect of peptide-centric versus spectrum-centric searching of DIA and DDA data. With spectrum-centric searching, the "currency" of detection are peptide-spectrum matches (PSMs), which are FDR corrected using target/decoy competition. With peptide-centric searching, the currency is a p-value for each peptide, where the FDR is estimated without competition. While it is possible to use spectrum-centric analysis for peptide detection, peptide-centric analysis is used for both DDA (MS1-level) and DIA (MS2-level). The re-use of ions in peptide-centric analysis has consequences over-reporting homologous or modified peptides, and we feel that the development of a hybrid analysis method by accounting for assigned ions in a peptide-centric search will be a necessary tool to ensure that FDRs are accurately assessed.

We planned three future tasks: to develop a perspective manuscript and two experiments. The perspective manuscript will discuss peptide-centric and spectrum-centric FDR, as well as the effects of shared evidence. We planned experiment for testing DDA versus DIA on the same sample to discover the sampling space for precursors and fragments. We also planned a second experiment for understanding target/decoy scoring for different decoy generation models using both synthetic and predicted target/decoy peptides.

4.9 Working Group Report: Proteomics Data Privacy

Juan Antonio Vizcaino (EBI – Hinxton, GB), Nuno Bandeira (University of California – San Diego, US), Eric Deutsch (Institute for Systems Biology – Seattle, US), and Benoît Kunath (University of Luxembourg, LU)

"Proteomics data privacy issues: Is proteomic data Personally Identifiable Information (PII)?" The detection of genomic variants at a proteome level implies that clinical sensitive proteomics data can be patient-identifiable, and then it should be protected appropriately (for instance, in the context of the GDPR (General Data Protection Regulation) guidelines in the European Union).

It is now the right time to assess the current state of the art and develop guidelines that are applicable to the community as a whole. Future data management policies for access to human proteomics data in the public domain are part of these efforts.

The main objective is to write a white paper that can be used to propose policy and to inform the community.

Participants

- Nuno Bandeira
University of California – San Diego, US
- Harald Barsnes
University of Bergen, NO
- Pedro Beltrao
EBI – Hinxton, GB
- Sebastian Böcker
Universität Jena, DE
- Robert Chalkley
University of California – San Francisco, US
- Lieven Clement
Ghent University, BE
- Frank Conlon
University of North Carolina – Chapel Hill, US
- David Creasy
Matrix Science Ltd. – London, GB
- Bernard Delanghe
Thermo Fisher GmbH – Bremen, DE
- Eric Deutsch
Institute for Systems Biology – Seattle, US
- Maarten Dhaenens
Ghent University, BE
- Joshua Elias
Chan Zuckerberg Biohub, US
- Michael Götze
ETH Zürich, CH
- Rebekah Gundry
University of Nebraska – Omaha, US
- Sicheng Hao
Northeastern University – Boston, US
- Nils Hoffmann
ISAS – Dortmund, DE
- Michael Hoopmann
Institute for Systems Biology – Seattle, US
- Lukas Käll
KTH Royal Institute of Technology – Solna, SE
- Michelle Kennedy
Princeton University, US
- Benoît Kunath
University of Luxembourg, LU
- Lennart Martens
Ghent University, BE
- Magnus Palmblad
Leiden University Medical Center, NL
- Hannes Röst
University of Toronto, CA
- Renee Salz
Radboud University Nijmegen, NL
- Birgit Schilling
Buck Institute – Novato, US
- Brian Searle
Institute for Systems Biology – Seattle, US
- Natalia Sizochenko
Dartmouth College – Hanover, US
- Stefan Tenzer
Universität Mainz, DE
- Yves Vandenbrouck
CEA – Grenoble, FR
- Hans Vissers
Waters Corporation – Wilmslow, GB
- Olga Vitek
Northeastern University – Boston, US
- Juan Antonio Vizcaino
EBI – Hinxton, GB
- Mathias Wilhelm
TU München, DE
- Bernd Wollscheid
ETH Zürich, CH
- Roman Zubarev
Karolinska Institute – Stockholm, SE