# Report of the SAB of the HUPO Human Proteome Project

to the Executive Committee of the HUPO-HPP

on the

HPP day session #2 entitled

“Moving the HPP Towards Proteome Functionality”

October 22, 2020

## 1. Executive Summary

On the occasion of its 10th Anniversary the HPP celebrated a string of major achievements. As a result of an international community effort a high stringency blueprint of the human proteome was published and an accompanying paper reports the credible detection of > 90% of predicted human proteins.  To the HPP-SAB this seems to be an opportune time to Initiate community discussion of major challenges facing proteomics and to develop plans for the HPP to address them over the next years. The SAB welcomed the opportunity to organize a session on occasion of the HPP day on Oct. 22nd 2020. This report summarizes the essence of the discussions and presents them as a contribution towards the planning of the next phase of the HPP.

## 2. Background and introduction

The international community effort of the HUPO HPP has resulted in a high stringency blueprint of the human proteome was published (1) and an accompanying paper reports the credible detection of > 90% of predicted human proteins (2). This represent a major milestone towards the goals set by the HPP at its inception, particularly by the C-HPP effort. Clearly, the efforts of the HPP in the near future will be directed towards the completion of the proteome map. Yet, the exploration of the (human) proteome and the elucidation of its biological significance, the central goal of the B/D-HPP, are by no means complete. Challenges that represent exciting opportunities for proteomics can be grouped into three areas.

The first concerns the complexity of the proteome. The HPP effort has resulted in the confident identification of one (sometimes several) particular translation products for most protein coding genes. In reality the number of protein species that constitute the proteome – proteoforms- is massively expanded compared to the number of protein coding genes by mechanisms including alternative splicing, protein processing and modification, and largely unexplored. Furthermore, proteins assume different conformations in 3D and associate with other biomolecules to form macromolecular complexes that often constitute functional units. Importantly, these latter properties of the proteome cannot be presently predicted from their sequence and recently a range of methods have become available to measure them.  The confident measurement of these properties of the proteome therefore provides an exclusive opportunity for proteomics to learn new biology.

The second area concerns the translation of molecular maps - the core proteome already mapped as well as forthcoming maps as described above- into biological function and phenotypes. It is generally well accepted that proteins either by themselves or in the form of modules are responsible for most biological functions. Knowing the proteins as chemical entities does not explain their biological function. The determination of the biological function of the proteoforms and modules that constitute the proteome is a major opportunity for the proteomics community. In that regard and to focus the task at hand, it is useful to consider different levels of function, including the biochemical and cell biological function of proteins.  The biochemical, catalytic function of a protein, exemplified by the phosphotransferase activity of a protein kinase or the proteolytic activity of a protease are of fundamental importance for the translation of molecular measurement into an understanding of molecular processes and mechanisms. The cell biological function of proteins considers their functional role in the context of the living cell. The cellular function of proteins needs to explain, e.g.

the set of proteins that are phosphorylated by a specific kinase in a particular cellular state and the impact of these events on the biochemical processes of the cell. The exploration of the cellular function of proteins and protein modules presents an outstanding opportunity for proteomics with large implications for basic and translational research.

The third area centers around the increased dissemination and uptake of proteomics techniques and knowledge resources as a mainstream component of life science research. It has been apparent for some time that proteomics uptake has been slow compared to e.g. genomic or imaging-based approaches, even though it is now widely acknowledged that proteomic data is particularly informative. Addressing the bottlenecks that prevent more rapid uptake and dissemination of proteomics presents a high yield opportunity for proteomics.

Session #2 of the HPP day in Stockholm was organized by the HPP SAB and focused on these three broad topics. Participants could join one of three parallel breakout sessions, each focused on one of the three themes. Rapporteurs led the discussion in each breakout session and then summarized the proceedings in a plenary discussion. The rapporteurs also agreed to summarize the discussions in writing as part of this report (see below).

## 3. Report from breakout sessions

### *3.1 Breakout session 1*: How can proteomics make the link from measuring molecules (proteins) to biological function?

Rapporteur Juri Rappsilber

Proteomics cannot:

- Provide mechanistic understanding => takes dedicated, focused and time intensive action by a specialist in the specific field or functional context of the proteins

Things proteomics can do:

- Direct analyses of protein behaviour
  - Quantitative measurement of perturbed system (readout is system-wide and could be proteins but also metabolites)
  - Individual gene knockout
  - Individual gene overexpression
  - Protein association to other molecules incl. RNA, DNA, metabolites
  - PTMs
- Functional association (not investigating the protein function itself but learning about the protein through its association with known proteins)
  - IP, Affinity pulldown
  - Proximity labelling
  - Co-localisation
  - Co-fractionation
  - Co-expression
  - Crosslinking

Emphasis was given to:

**Readout**. Essentiality (under the conditions studied) is easy to detect through a gene knock-out. But impact on the proteome is also a measurable readout these days thanks to advances in proteomics. What constitutes a scalable readout will be influenced by technological progress.

**System and condition**. Biological variability, inter-person variability, different cell types (over 200), diseases (e.g. viruses rewire the cell completely), influence of laboratory conditions all impact to various degrees on the readout and should be considered and included.

**Openness to other fields**. Proteomics is not the only technology that can add towards the functional understanding of proteins.

**Fundability**. This might require lobbying. This might also link to usefulness to others, which raises the question of what would be useful to others. In any case, if the initiative is not designed such that individual labs can acquire funding then the magnitude of it will be very much limited.

### 3.2 Breakout session 2: How can HPP explore the full extent of the complexities of the proteome and translate knowledge into new biology

Rapporteur: **Kathryn Lilley**

At the HPP meeting of HUPOconnect2020, a breakout session considered how to move the identified proteins now mapped to most open-reading frames from a simple list to a resource representing the functional complexity of the proteome. This could be viewed as akin to currently viewing a 'photograph' of the human proteome compared to a 'movie' of moving parts with temporal, spatial and physico-chemical definition. To fully access the functional proteome, we need to characterise each possible proteoform that can occur through post transcriptional and post translational modification in an infinite number of biological settings. To achieve this aim, we will be required to collect a multitude of different datasets reporting on activity status, interacting partners (including other proteins, nucleic acids, lipids and metabolites), subcellular location, tissue specificity and expression levels, and heterogeneity within single cell populations. It was recognised by this discussion group that with current available technology, this is an impossible task at the scale required. There is a clear need for transformative technology development and application to reach these aims.

To put the above into a some sort of numerical context, our understanding of the chemical entities that form the proteome is only the tip of the iceberg as there are an estimated 70000 alternatively spliced transcripts and each protein can be heavily modified after translation, leading to an estimated 2.8247525e+48 different chemical protein entities or proteoforms (making an assumption that an average of 10 combinatorial PTMs per protein can occur per protein).

Each proteoform may have a unique function and we discussed at length how the transcriptome is not giving us many clues about alternative functionality amongst related proteoforms.

We also acknowledged that measurements of the proteome are very biased towards proteins expressed in most tissues and cell types and that fall within a set of physcio-chemical parameters that suit the protein extraction and manipulation methods commonly applied.

A concern we had was that the field of proteomics is viewed somewhat simplistically by the community, and that the 19733 human proteins currently mapped to ORFs are pretty much the human proteome. The sheer number of potential proteoforms and their varied context specific functionality is accepted by the proteomics community, but not necessarily by the community at large. This may lead to a lack of understanding of proteomics data sets and erroneous interpretation of what data are showing – i.e. all proteoforms of an identified protein being lumped together as a single observation.

**We discussed that making the community aware of the significant shortfall in proteoform level analyses is of paramount importance and is timely.**

We also discussed how to go about starting to collect the breadth of data needed to fill in our gaps in our understanding of the functional proteome. We acknowledged the fact that existing data residing in repositories could be further mined, as many deposited datasets remain under-exploited.
In terms of new data collection, it was recognised that funding could be an issue as there is no budget from HUPO to for a project of this scale. Any project also would need to be approached in a systematic manner, possibly focussing on one disease state, such as malaria or cardio vascular disease. Structured HPP projects, taking one aspect of a system, such as post translational modifications or protein protein interactions, could also be a way forward, with teams of researchers from across the globe coming together to work on a focussed project. The need to capture data and create an efficient web portal. It was suggested that there could be 'ballot' set up for such projects similar to how HUGO conducted its research in the past.

In short, the enormity of the task in hand to fully understand the functional proteome cannot be underestimated. It will take, a lot of organisation and resources and can only be approached as a global team effort. How to attract the considerable funds required other than piecemeal for different nation and cross national funding streams is unclear. How to come up with a blueprint for a project of such magnitude also remains unclear, but a period of intense 'naval gazing' is required to scope out such an endeavour and plan series of pilot experiments that will inform interrogation of the functional proteome at scale.

### *3.3 Breakout session 3:* What are the bottlenecks to increase the uptake and dissemination of proteomics?

Rapporteur: **Cathy Costello** (Text extracted from session recording)

The breakout group identified a number of factors that have limited the uptake of proteomics, particularly in the translational field and identified a number of measures that could be implemented to improve the situation.

Challenges:

- Cost and complexity of mass spectrometers. Whereas instrumentation is still complex, the field has reached a level where the performance of lower cost instruments is well advanced and the routine generation of high quality, clinically relevant data is feasible.
- Availability of trained personnel at different levels of education.
- Access to proteomic technologies and sample to data turnaround time.
- Access/dissemination of present state of the art and capabilities to clinical scientists.

Proposed measures:

- Train new researchers in the proper operation of high and lower level instruments
- Develop mechanisms for retention of trained personnel. Make proteomics into an exciting career track for the best scientists.
- Obtain funding.

- Put more emphasis on patient advocacy. This is very effective. Identify patient advocates who have benefitted from proteomics. Invite some patient advocates to meetings. And engage them to prepare proposals for funding.

- Look within the B/D activities for those where proteomics has an effect on the disease and then publicize the cases

- Look to industrial partners to publicize what can be done

- Identify areas where proteomics could be particularly effective.

- Establish links with existing initiatives, e.g. Perdita Berran's program

- Document some specific cases of people and institutions as examples of exciting proteomics programs.

## 4.Next steps

This is an incomplete draft for initial discussion by the HPP EC. The SAB will complete the document as soon as possible and then formally submit it to the EC. It is hoped that the report and the ideas it contains can serve as a basis for discussion towards the next phase goals of the HPP. The timing of these discussions, the format and location has to be determined by the HPP EC.

## 5.Acknowledgements

Thanks go to Rapporteurs, Kathryn Lilley, Cathy Costello and Juri Rappsilber, the HPP EC for their support of the session and to all participants and discussants.

## 6. References

1)Adhikari S, Nice EC, Deutsch EW, Lane L, Omenn GS, Pennington SR, Paik YK, Overall CM, Corrales FJ, Cristea IM, Van Eyk JE, Uhlén M, Lindskog C, Chan DW, Bairoch A, Waddington JC, Justice JL, LaBaer J, Rodriguez H, He F, Kostrzewa M, Ping P, Gundry RL, Stewart P, Srivastava S, Srivastava S, Nogueira FCS, Domont GB, Vandenbrouck Y, Lam MPY, Wennersten S, Vizcaino JA, Wilkins M, Schwenk JM, Lundberg E, Bandeira N, Marko-Varga G, Weintraub ST, Pineau C, Kusebauch U, Moritz RL, Ahn SB, Palmblad M, Snyder MP, Aebersold R, Baker MS. A high-stringency blueprint of the human proteome. Nat Commun. 2020 Oct 16;11(1):5301. doi: 10.1038/s41467-020-19045-9.
2) Omenn GS, Lane L, Overall CM, Cristea IM, Corrales FJ, Lindskog C, Paik YK, Van Eyk JE, Liu S, Pennington SR, Snyder MP, Baker MS, Bandeira N, Aebersold R, Moritz RL, Deutsch EW. Research on the Human Proteome Reaches a Major Milestone: >90% of Predicted Human Proteins Now Credibly Detected, According to the HUPO Human Proteome Project.
J Proteome Res. 2020 Dec 4;19(12):4735-4746. doi: 10.1021/acs.jproteome.0c00485. Epub 2020

Zürich 2-02-2021, The HPP SAB:

Ruedi Aebersold, (chair)
Subhra Chakraborty
Anne-Claude Gingras
Fuchu He
Kathryn Lilley
Emma Lundberg
Anthony Purcell
John Yates