MARCH 2022

ADVANCES IN PROTEOGENOMICS:

HOW PROTEOMICS CAN COMPLEMENT GENOMIC ANALYSES

A **GENOMEWEB** VIRTUAL ROUNDTABLE DISCUSSION

genomeweb

SPONSORED BY:

Olink

Given that proteins are the primary functional molecules of cells, many researchers are interested in studying genomic data in combination with proteomic information. And, for good reason: Such proteogenomic analyses can empower scientists to better determine the functional impact of genetic mutations discovered by next-generation sequencing or identify new potential drug targets.

The possibilities are significant but so are the challenges. To provide the scientific community with insight into how to optimally leverage proteogenomics, four leading experts – Michael Snyder, professor of genetics at Stanford University; Claudia Langenberg, professor of computational medicine at the Berlin Institute of Health at Charite; Karin Rodland, affiliate professor of cell, developmental, and cancer biology at Oregon Health and Science University and laboratory fellow at Pacific Northwest National Laboratory; and Janne Lehtiö, professor of medical proteomics at Karolinska Institute and director of the Clinical Proteomics Mass Spectrometry Facility at SciLife Lab – shared their insights during a GenomeWeb Virtual Roundtable, sponsored by Olink. The presentation was followed by a question-and-answer session with the audience. This report is based on the online discussion, which was moderated by Adam Bonislawski, senior editor at GenomeWeb.

EMBRACING AN INNOVATIVE APPROACH

The panelists started by zeroing in on the most basic question: **Why should researchers adopt proteogenomics?** Their consensus: While proteomics and genomics offer numerous advantages as standalone disciplines, researchers get the best of both worlds when the two fields of study are combined.

Overall, proteogenomics supports a framework that enables researchers to discover how proteins reflect health and disease states. More specifically, proteogenomics empowers researchers to study genetic determinants by providing insights on the causes of protein levels and variants, which results in a better understanding of disease mechanisms. By doing so, this emerging discipline can help with disease prediction and prognosis, Langenberg said.

MODERATOR:

ADAM BONISLAWSKI

Senior Editor
GenomeWeb

PANELIST:

CLAUDIA LANGENBERG, PhD

Professor,
Computational Medicine
Berlin Institute of Health
at Charite

PANELIST:

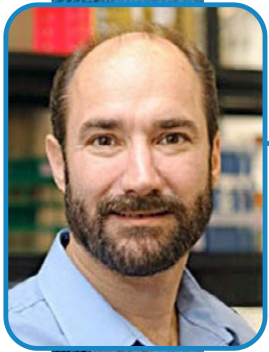
JANNE LEHTIÖ, PhD

Professor of Medical Proteomics
Karolinska Institute
Director, Clinical Proteomics
Mass Spectrometry Facility
SciLife Lab

PANELIST:

KARIN RODLAND, PhD

Affiliate Professor, Cell,
Developmental, and
Cancer, Biology
Oregon Health
& Science University
Laboratory Fellow
Pacific Northwest
National Laboratory



PANELIST:

MICHAEL SNYDER, PhD

Professor, Genetics
Stanford University

Snyder agreed that such insights result in a better read-out of a phenotype than what is produced from RNA studies, which do not enable researchers to assess protein levels. As a result, proteogenomics can more effectively predict disease states from a biomarker standpoint, he said.

Rodland continued the conversation by pointing out that the most significant advantage is found in post-translational modifications, which are invisible when relying on DNA and RNA. "We can actually measure changes in protein phosphorylation and determine with great accuracy which pathways are responsible for mediating the response at the RNA level," Rodland said. "This is extremely important if we want to identify therapeutic targets."

Rodland provided an example in which she and colleagues leveraged analyses of phosphorylation of substrates to identify that Aurora kinase B is the most highly active kinase in the early phases of developing resistance to Fetal Liver Tyrosine Kinase 3 (Flt3) inhibitors in acute myeloid leukemia (AML). The findings, the team wrote, can inform new targeted AML treatment strategies. "This is a nice complete story where, in addition to the RNA-seq and the genome data, adding in the phosphoproteomics gave you an additional target that you would not have gotten from genomics alone," Rodland noted.

EXPLORING POSSIBILITIES

Because proteogenomics combines different data types in one location, it helps researchers maximize the power and versatility of scientific inquiries.

The pragmatic uses are many, as proteogenomics can be used "across the board" to study a variety of conditions including "common complex diseases, cancers, infectious diseases, and COVID prognosis," Langenberg said.

Researchers can combine knowledge of the genetic signal of a protein-encoding gene with information about the genetic signal for specific diseases. Adding information about expression levels makes such research even more valuable, Langenberg added.

In addition, with proteogenomics, researchers can rely on publicly available data to establish new links without necessarily studying each disease again from the start. "Hence, one of the most amazing uses has been that we can use this not just to understand how the genetic signals for the protein may be shared with a specific disease, but how that is shared with … other diseases that seem to be unrelated etiologically," Langenberg said. Ultimately, such insights lead to a better understanding of which mechanisms connect different diseases, as well as to the development of a gene-centric definition of disease nomenclature.

Snyder added that proteogenomics brings proteomics and metabolomics data together to better solve "mystery diseases." In addition, he noted that proteogenomics can help "to predict disease risks from personal genomes. So, looking at variants and seeing which ones are expressed and how they affect the levels of proteins is powerful," Snyder said.

What's more, proteogenomics can support cancer neoantigen discovery. With proteogenomics, it is possible to progress backward from peptide to cancer genome data and uncover unexpected coding regions. Doing this in human samples enables researchers to "find novel coding regions and novel proteins by resurrecting pseudogenes that [can be found] in a normal tissue as well. So that adds a new type of data that is very difficult to generate by DNA and RNA," Lehtiö said.

Pacific Northwest National Laboratory, like many others, is leveraging proteogenomic pathway discovery to expand its research efforts. "There's been a lot of discussion about neoantigens and the ability to identify those and use them in a vaccine approach to cancer. However, it's not just the neoantigens, it's the antigen-presenting machinery as well that determines the effectiveness of an immune response to the neoantigens," Rodland said.

In addition, the laboratory conducted Clinical Proteomic Tumor Analysis Consortium (CPTAC) studies that are directly applicable to cancer therapy. The studies show that the tumor mutational burden (TMB), which is often used as a biomarker of response to immunotherapy, does not correlate either with the number of neoantigens or the efficiency of the antigen-presenting machinery, Rodland noted.

Similarly, the Clinical Proteomics Mass Spectrometry Facility at SciLife Lab is using proteogenomics to address lung cancer. With this approach, researchers can assess a comprehensive view of the total neoantigen burden, including both the TMB that comes from the traditional genomics point of view as well as the mutational and neoantigen burdens. As such, researchers "can see lots of complex neoantigens, where long stretches of amino acids are produced by exon/intron boundary mishaps and other types of events," Lehtiö explained.

The lab found that known immune evasion mechanisms are related to point mutations. By examining different immune evasion mechanisms, it also discovered a soluble immune inhibitory protein that commonly secretes in tumors with a complex neoantigen. The genomics data made it possible to link the neoantigen to the serine/threonine kinase 11 (STK11) mutation that activates liver-specific transcriptional regulation of the immune evasion mechanism, illustrating how the neoantigen analysis, proteomics, and genomics can work together. This proteogenomic data empowers researchers to "start seeing new hallmarks that single mutations can activate or deactivate," Lehtiö added.

CONFRONTING CHALLENGES

While researchers are discovering the power of proteogenomics, they simultaneously need to clear a variety of hurdles such as:

Capturing the diversity of the proteome:
In proteogenomic studies, proteins in a sample are broken enzymatically into peptides, separated with liquid chromatography, and then analyzed with mass spectrometry. However, the liquid chromatography step, Lehtiö said, can act as a sort of bottleneck, excluding certain peptides from downstream analysis. Peptides that vary by a single amino acid or by a small post-translational modification may not be separated by liquid chromatography and are then conflated during mass spectrometry. This means researchers must sometimes revert to the gene-centric view of the proteome, which does not capture the diversity of modifications and alternate splicings of the proteome.

Similarly, while the peptide fragments analyzed by mass spectrometry can usually be accurately matched to predicted proteins, the error rate increases when searching among greater datasets of proteins. "When you start using the heaps and heaps of genomics data, the database size increases. That lowers your sensitivity or increases the error rate. So that's one of the things that has been problematic in feeding the sequence variants," Lehtiö said. Accurately determining point mutations also "can really derail in sequence variant analysis," Lehtiö added.

However, Lehtiö said, technological advances in chromatography and informatics are helping to solve that problem. "About 12 years ago, the pressure in the [high-performance liquid chromatography] systems was 300, and now we are up to 1,500 bars. That has really sharpened the chromatography. At the same time, we have learned lots of bioinformatics tricks to segment the databases and [improve] the data analysis," Lehtiö said.

Throughput:
"Historically, the throughput of proteomics has been a lot lower" than what is achieved in genomics and transcriptomics, said Snyder. However, with scalable capture technologies, such as those from Olink and Somalogic, throughput is improving, he noted.

Quantification of analytes:
"Getting absolute quantification has been a challenge. We're better at it now than we used to be, but we still have some improvements to go," Snyder said.

Cost:
While proteogenomics costs are decreasing, research organizations are still looking for increased affordability. "They're not $50 an assay like they should be. When that happens, everybody will be doing much more," Snyder said.

Achieving single-cell proteomics:
Studying tumor cells individually to understand why some become resistant is a challenge that could be made more manageable by emerging technologies. "We're on the cusp. Some technologies are coming out, especially in the spatial place and a little bit in single-cell profiling," Snyder said.

FORECASTING FUTURE DEVELOPMENTS

In addition to overcoming specific challenges, the panelists noted that scientists are pushing proteogenomics forward by:

- Using chromatography and fractionation to segment databases and develop the mass spectrometers.
- Taking advantage of new affinity-based proteomics methods to uniquely marry discovery and high-throughput validation.
- Leveraging complementary affinity- and mass-spec-based methods to perform either a large- or small-scale proteomics analysis.
- Utilizing broad capture technologies to enhance disease prediction by making it possible to discover the value of proteins that were not previously measurable.

In addition, the future is ripe with opportunities to advance scientific inquiry. For example, the ability to leverage single-cell proteomics or to determine spatial resolution on diameters that are relevant to tissue proteomics is set to become a more common reality. "Partial proteomics on the order of 100 square microns is now really quite feasible, and we can get almost as good coverage at the unmodified protein level as with much larger samples," Langenberg said. "We're going to see, at the very least, histologically relevant spatial proteomics within the next two or three years. And I'm optimistic that we will see true single-cell proteomics within the next five years."

Increased interaction among metabolomics, proteomics, and genomics as well as RNA studies is also an expected development. By using these together, scientists can support mechanistic research. For instance, they have already discovered how cholesterol is reduced by fibers, which would have been difficult to do with proteomics alone. "But together it looks like one of the major mechanisms for fiber removing cholesterol … is through bile acids. It took all the different technologies to [determine] that," Snyder said. "That's one example of what will be many in the future, where the combination will become so routine."

Genome sequencing also figures to play prominently in the days ahead. Researchers are "going to want to know what that looks like in terms of phenotypes, and proteomics is going to be one of those assays. [Researchers will] want to know what's going on with your variants and other peoples' variants, especially if they have a mystery disease," Snyder said.

Population health and clinical applications bubble to the top when Langenberg considers the future of proteogenomics. The data, which will be widely available, will come with a variety of uses and will support the "mechanistic insight and investigative ability to predict a whole range of different diseases that are currently very poorly identified," Langenberg said. "Patient-based studies that can look at prognostic stratification and differences in prognosis, and that [applies to single-cell and also to block-based] biomarker studies, is a huge area that's been largely untapped."

Lehtiö agreed that the most promising future developments are apt to focus on clinical applications. "We are very interested in using proteogenomics in clinical trials, and incorporating the molecular phenotype with genotype and clinical phenotype. ... That's very important," Lehtiö said. In addition, researchers have leveraged cancer genomics to connect one mutation to one drug but "that's not curing the cancer patients. There is relapse quite often. So, we need to combine multiple drugs to actually hit several cancer hallmarks," Lehtiö added.

Proteogenomics also can help scientists delve deeper into cancer immunotherapies. "The whole immune system with all its mediators, soluble mediators, and receptors, is covered by proteomics and proteome. So, we really need to look at that in order to understand the immune evasion mechanisms and the early responses," Lehtiö concluded. "Proteogenomics is going to be very valuable in all aspects: in diagnostics, in stratification, and prognostics, but, very importantly, in precision medicine and monitoring early treatment response."

AUDIENCE Q&A

The following question-and-answer session has been lightly edited for clarity and length.

Adam Bonislawski:
What is the status of doing this at the single-cell level and bringing some sort of spatial information to bear on these questions? And what could that possibly provide in terms of new insights?

Michael Snyder:
We actually do a lot of CODEX, which, many of you know, stains with about 50-some-odd antibody probes. You've got to get those probes validated first. But then you can really collect a lot of information.

What's special about spatial is you not only see what's there, but you actually see the arrangement. The goal these days is to try to understand neighborhoods. So, for example, cancer cell neighborhoods are very different from normal tissue neighborhoods. And it's really quite cool. We have a project, for example, where we're looking across the intestine, both small and large, at immune cells, and you see how the neighborhoods change. It's quite fascinating actually.

So, I think that's going to be one of the next futures, if you will, of the space, is to try and understand not just what cells are there, which is what we do now. But really trying to understand their neighborhood and how that changes cell states and cell behaviors. Presumably, it's through cellular interactions. You can capture all that with the appropriate probes. So, CODEX is certainly one technology. There are other technologies that are emerging now, it's a very, very competitive space.

I saw on the chat somebody was asking about subcellular localization. You can get that from CODEX. I don't think most people take advantage of that. But there are other systems like Resolve Biosciences, they have a molecular cartography system, mostly around nucleic acids that'll do subcellular as well. So we're heading that way, not just cellular, but subcellular as well. So, get ready.

Janne Lehtiö:
I can say a few comments on the mass spec side on the single cell. It's a bit exaggerated to say that we can do mass-spec single-cell proteomics. But the technologies really are going forward quite fast. But I think that we are a bit far away. But I think the multiplexing in separated cells to look at the surface proteome is also very interesting, both with a sort of multiplexer [fluorescence-activated cell sorting] type of analysis and, again, the mass cytometry-based analysis.

One thing that is important, I think, is to relate the bulk data to the spatial regulation in these multiplex imaging things, to look at the proteins in their neighborhoods and context like Mike said, and also the single-cell situation. So I think that that's a very, very interesting development that is going on and is also marrying mass spec with affinity proteomics quite nicely.

Adam Bonislawski:
Can proteomics catch up with RNA-seq in terms of throughput and coverage? Can that gap be closed? And what needs to be done to get there?

Karin Rodland:
In what context? We have to define the context in which we're trying to do the comparison. If you're talking about ample tissue, like in the [Clinical Proteomic Tumor Analysis Consortium] studies, then we're already there, I think. If you have ample material, you can get as in-depth coverage with proteomics as you can with RNA-seq, and they can be very complementary to each other.

The problem is that RNA-seq is an amplification-based technology and proteomics is not. So that when you get to smaller sample sizes, like spatial and single-cell, proteomics cannot catch up to RNA-seq. In the absence of some huge technological breakthrough that compensates for the inability to amplify proteins, I don't see that happening.

But, again, you're asking different questions of the RNA-seq data and of the protein data and you're asking different experiments. So, if the point of the experiment is to catalog every transcript and every protein product in the cell and look at the relationship between the presence of a transcript, a micro-RNA, a long non-coding RNA, and the abundance of the cognate protein, we are there. We can answer that question.

Michael Snyder:
I think we're almost there. I still think the throughput and the cost are a little bit cheaper on RNA-seq than for proteomics. Our lab does both and we can just do a thousand RNA-seq experiments pretty easily in a week. It's so hard to do that right now with proteomics. But I think we can get there and I think some of the technologies, both with mass spec and – again, I think this is where the capture reagents are really quite powerful for being able to increase throughput. Somebody put in the chat a question about the deCODE paper, I think that was a SomaScan paper. But they did 35,000 samples. That was not possible before.

So we are getting there, but it's pretty expensive to do those compared to RNA-seq. But what's going to help a lot is that there's just been a ton of interest, and to be honest, the investments are now increasing in proteomics. Some of you may have heard me say this before. But Seer went public for a $3 billion valuation, and that adds a lot of interest to this area, which I think is going to stimulate, I hope, a lot of activity, which in turn should drive down the cost and increase the throughput.

Then I think we will be there and I think people will prefer proteomics data over RNA-seq data because, again, it's close to the phenotype as talked about much earlier. So that's my opinion, but I recognize there are other opinions out there.

Adam Bonislawski:
What role do newer technologies, such as protein sequencing-based technologies – like those from Nautilus or Quantum-Si, or even pre-commercial things like nanopore sequencing – potentially have to play in proteogenomics? And are these going to be able to do things that we can't do right now?

Michael Snyder:
I actually think there's a huge potential here. It's very early days. It kind of reminds me of some of the nucleic acid early-day technologies where they're expensive, they're clumsy. But I'll make a plug for the [US Human Proteome Organization] meeting coming up, I know some of the vendors are going to be there, and I'm going to definitely hit their booths because I want to see what the latest is.

I think the ability to be able to display all proteins and probe them or sequence them has the potential for incredibly high throughput and incredible sensitivity for doing single-cell proteomes in a very orthogonal way. And I know there's a lot of other technologies out there, some well-advertised and some less well-advertised, that really could be game-changers. There's a number of groups all working on, for example, pore technology for trying to sequence, and I imagine a nanopore equivalent for proteomics would be quite powerful.

So I think there's a lot of that stuff in the research phase. The Nautilus stuff intrigues me for drug probing purposes and things like that. So there's just a lot of very cool stuff out there. As they mature, I think they will find a way into the mainstream and give us capabilities we're currently not doing so.

Janne Lehtiö:
I think Mike is completely spot-on because there are a few things in those technologies that are very interesting to keep an eye on. That sort of capability of single-molecule detection and a read-out of single molecules could really be a breakthrough. You could imagine that if you could read longer stretches of protein sequence, we could solve some of the protein inference problems that we have matching the peptide data and variant data, and so on. So there are lots of struggles still there.

But the other thing that is remarkable with some of these possibilities is the way that you can massively parallelize the analyses. So you could actually get single-cell detection in a massively parallel way, and that could really be a new view in the proteome. But I mean, it will need some heavy investments and lots of technological breakthroughs.

Michael Snyder:
You know what it reminds me of, Janne? It reminds me of the Oxford Nanopore days, where they had this buzz going literally for years. Every year we're going to see the latest coming out from Oxford Nanopore. And it

struggled and it struggled. But it ultimately did come out and now it's a pretty powerful technology.

That's what a lot of these things are reminding me of. I think it's very exciting. But it may take time to mature, especially if they're going to commercialize. Nobody wants to pay for something that doesn't work.

Janne Lehtiö:
I was also doing my PhD in a lab next to the one that invented pyrosequencing, sequencing by synthesis. For years it looked completely hopeless, and I thought, the poor guy, it's never going to work. Then all of a sudden they just solved some technical issues and really kicked off. So, you're right, that you never know. We should keep eye on that.

Adam Bonislawski:
How do you translate this information into clinical data? Where do things stand in making that translation?

Claudia Langenberg:
It's a huge privilege to end up with results that kind of scream at you. But if they happen across multiple dimensions and happen at that scale, it's a challenge. So I think I can only talk about a few personal experiences that have made it a lot easier for us. The team has to have very different disciplines.

I also think more efforts should be made in visualizing the results better. We have gone beyond massive Excel spreadsheets that we could sit at and read. So we need to create innovative ways of integrating the different biological domains and bring them to our results from proteogenomics. So I think that has been one of the challenges. But it has also been one of the fun parts because people really then engage very differently with the data, particularly if it's interactive. It can help biological interpretation by making the various resources that are now thankfully available much more accessible and bringing them together in one place.

Then, finally, nobody is an expert across so many different specialties, and hence, to think about how you get clinical engagement from people who have not thought about this protein ever, obviously, that's another task. So that requires more communication, and it may be a very simple message. But, how do we manage to engage the relevant clinical experts to take what we have and interpret it in the relevant clinical context? So that's the other thing that I think is very important.

Karin Rodland:
I'm seeing more and more uptake of the addition of proteomics into precision medicine clinical trials. Consortia like the Clinical Proteomic Tumor Analysis Consortium had certainly helped with this, the Beat AML Consortium is now adding proteomics to their very large-scale multi-institutional analysis of patients with AML.

As more of the precision medicine studies are adding proteomics and getting new information, just like Claudia said, that directs therapy or directs prognosis and risk stratification, they're seeing the value of adding the proteomic component. So there's a greater and greater adoption.

Now, one of the barriers to the field or just where we are right now is that there's skepticism about ever being able to get CLIA certification for a mass spec-based assay or an LC-MS-based assay because of technical issues. So the pipeline still appears to be that one uses proteogenomic approaches in a discovery mode to look broadly, as Claudia said, and find that new interaction, that new relationship, that new prognostic factor that hasn't been spotlighted before. Then one goes on to targeted assays that may be mass spec-based, like PRM or SRM. And then there's still a feeling that we need to go to an ELISA-type assay for clinical use.

But I see a day when I think that the mass spec will be in the clinic. There are already studies, and Matthew Ellis has done this, where, from a needle biopsy, you can get DNA, RNA, denatured protein, and native protein, and you can get enough of those to send them off for targeted assays, looking for specific markers of prognosis or treatment. So I do see that day coming. ■