# Detecting Scarce Isoform-specific Peptides Using Next-Gen Protein Sequencing™ (NGPS™)

**Gloria Sheynkman, PhD**

Assistant Professor, Department of Molecular Physiology and Biological Physics
University of Virginia School of Medicine

***"The proteome is chemically and functionally complex and multidimensional, and the analytical techniques used to characterize it, such as mass spectrometry, can only get us so far. There is a ceiling. What attracted us to Next-Gen Protein Sequencing is the single-molecule detection technology. This allows us to detect proteoforms with subtle amino acid variations – small changes that can have big implications for protein function."***

– Gloria Sheynkman, PhD

**Challenge:** Detecting proteoform variations is challenging: current tools face issues with sensitivity, throughput, sample preparation, and the ability to distinguish between proteoforms that have similar mass and charge

**Innovation:** The use of long-read RNA sequencing, advanced analytical and computational approaches to predict full-length protein isoforms, and understanding their mechanisms to drive disease states

**NGPS Integration:** Quantum-Si's Platinum® instrument detects proteoforms at the single-molecule level, complements mass spectrometry by accessing undetectable proteome regions, and resolves isobaric peptides using N-terminal amino acid recognizers

**Advantages:** Single-molecule sensitivity identifies rare proteoforms that mass spectrometry may miss and distinguishes subtle sequence variations, such as leucine vs isoleucine, at a lower cost and with less training required

Q398 REV 01

The number of proteins expressed by our 20,000 genes may total 300,000 or more. Distinct protein isoforms, called proteoforms, result from post-transcriptional mechanisms, such as alternative RNA splicing, and post-translational modifications (PTMs) including phosphorylation and proteolytic cleavage. While these mechanisms produce physiologically normal proteoforms, disruptions in the process result in the expression of aberrant proteoforms that can lead to a wide range of diseases. For example, a recent top-down proteomics study evaluating KRAS4A and KRAS4B (isoforms of KRAS, one of the most frequently mutated oncogenes in human cancer), revealed 39 aberrant proteoforms across colorectal cancer cell lines and primary tumor samples.[1] Similarly, a top-down proteomics study demonstrated that the tumor suppressor protein PTEN has been shown to exist in a number of different proteoforms.[2] Given the role of such protein variants in disease, there is an urgent need to identify disease-associated proteoforms to enable development of proteoform-based biomarkers and therapies.

Dr. Gloria Sheynkman, PhD, and her team at the University of Virginia are using advanced analytical and computational approaches to investigate how proteoform variation underlies human disease. They are developing approaches to discover novel disease proteoforms, assay proteoform-specific functions, and elucidate the molecular mechanisms by which proteoforms rewire cellular networks to drive disease states.

We spoke with Dr. Sheynkman about how Next-Gen Protein Sequencing (NGPS™) on the Quantum-Si Platinum® instrument serves as a powerful and complementary technology, enabling access to and interrogation of the full proteome.

**Q: What is top-down proteomics?**

**A:** Top-down proteomics is the analysis of intact proteins by mass spectrometry. For many years, it has been the only technique in the game for characterizing complete proteoforms that arise from alternative splicing events and PTMs. Among the benefits of this approach is that it allows for exact mass calculations of proteins and does not require digestion of proteins into peptides. In contrast, the bottom-up method does require peptides as the starting point and this can lead to the loss of proteoform information as many homologous sequence regions are shared between different protein forms.

**Q: What challenges are associated with top-down proteomics?**

**A:** There are some well-recognized technological challenges including sensitivity, throughput, and sample prep. If we are trying to define the entire proteome using mass spec, we face two major obstacles. The first is the difficulty of distinguishing between protein molecules that are very similar in physicochemical properties such as mass and charge. The second is the difficulty of measuring the proteome in general, which is difficult because the dynamic range of protein concentrations can span ten orders of magnitude in biological samples such as serum.

I think another obstacle is cultural in its origin. It's related to how scientists design their experiments and how hypotheses are formed and that has been driven, historically, by the tools that are available. If you don't have the right tools and you cannot easily measure proteoforms, then you're not thinking about how biological systems are driven by individual proteoforms.

Related to this, conceptually, is that we have gene ontology but not proteoform ontology. Most functional annotations are organized around genes, which has served us well. We're capturing diseases where the genetic evidence is clear. But we're missing a huge sector of biology in many diseases where the evidence is less certain or non-existent. This includes non-coding variants and how they affect the proteome and biological systems downstream. It's like a dark region.

**Q: Talk a bit more about bottom-up proteomics.**

**A:** Top-down proteomics was the next logical step from bottom-up proteomics. While its much more analytically tractable to measure peptides versus intact proteins, with peptides you lose the provenance. You might detect a peptide that tells you that, at a certain site, a variation exists, but you don't know how it is connected to other variations of the same protein; in other words, you don't know from what exact proteoform a peptide arose. There's a lot of missing information and we ultimately want the full atomic structure of the proteoform because even the smallest change can lead to significant effects in terms of health and disease.

**Q: How are you integrating Next-Gen Protein Sequencing on the Quantum-Si Platinum instrument into your research?**

**A:** We use long-read RNA sequencing to derive full-length, fully resolved transcriptomes from sample cell lines and tissues, and we use that information to predict the possible proteins that could be expressed. Such highly personalized protein references will be critical for personalized medicine, and studies across the full human population across ancestries. We have been using bottom-up proteomics, but mass spectrometry can only get us so far. There is part of the proteome that will not be detectable via mass spec.

That's why we're excited about Platinum, as we needed a complementary but completely different, powerful technology to uncover blind spots of MS so as to interrogate the full proteome. Being trained as an analytical chemist, I feel that the more orthogonal you can get, the better, and this is true of Next-Gen Protein Sequencing and mass spec.

The other thing that attracted us to the Quantum-Si platform is the single-molecule detection technology, which means it is possible to detect proteoforms with subtle amino acid variations that can have huge implications for their functions.

**Q: Tell me about the process to get up and running on the Quantum-Si platform.**

**A:** Mass spectrometers can cost more than a million dollars and typically require a PhD-trained senior expert to run them. With the Quantum-Si platform, we bought it in a week, had it installed, and our second-year grad student started running it after a few days of training. It also takes up very little space on the benchtop – it's like the size of a thermal cycler. The bar is much lower to get started compared to mass spec.

**Q: What are your initial impressions of the platform?**

**A:** One of the things that we have seen so far relates to isobaric peptides which have very close physical and chemical properties; an example would be a leucine to an isoleucine switch. Its difficult to distinguish between those using a typical liquid chromatography-mass spec as they would elute at the same time, and their mass-to-charge ratio is identical. It's ambiguous. Because the Quantum-Si platform uses N-terminal amino acid recognizers to sequentially identify amino acids in a peptide, we can distinguish peptides that have a variation of leucine versus isoleucine on a single-amino acid level.

Another advantage arises from the fact that many proteoforms consist of small, subtle changes that all come together to make a whole new functional protein. If we can detect a canonical isoform versus a minor isoform that has implications in disease or complex traits, and distinguish such molecules at the single-amino acid level, that is very exciting for our team.

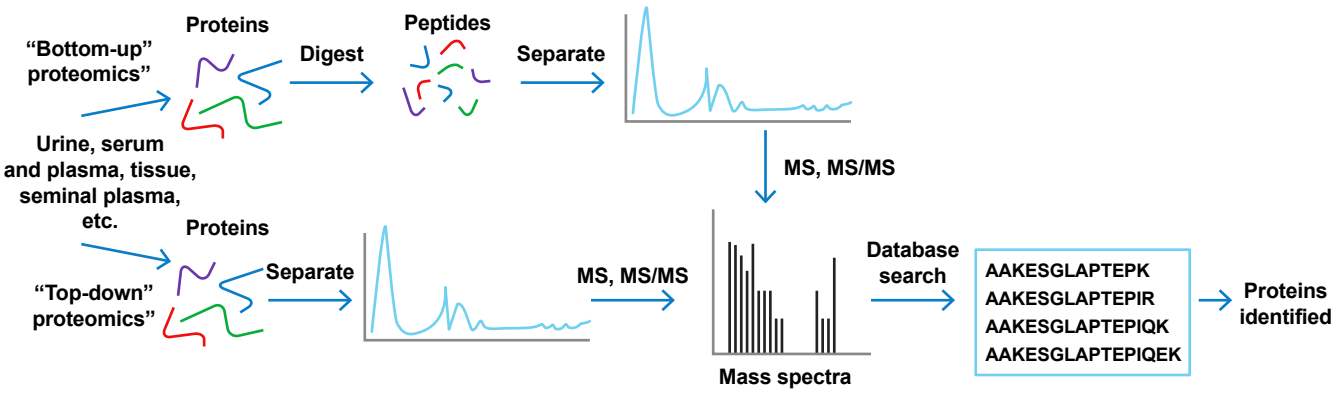


**Figure 1.** Workflow for bottom-up and top-down proteomics.[3]

## REFERENCES

1. Adams LM, DeHart CJ, Drown BS, Anderson LC, Bocik W, Boja ES, et al. Mapping the KRAS proteoform landscape in colorectal cancer identifies truncated KRAS4B that decreases MAPK signaling. *J Biol Chem*. 2023;299(1):102768. doi:10.1016/j.jbc.2022.102768.
2. Malaney P, Uversky VN, Davé V. PTEN proteoforms in biology and disease. *Cell Mol Life Sci*. 2017;74(15):2783-2794. doi:10.1007/s00018-017-2500-6. PMID: 28289760.
3. Meo A, Pasic M, Yousef GM. Proteomics and peptidomics: moving toward precision medicine in urological malignancies. *Oncotarget*. 2016;7. doi:10.18632/oncotarget.8931.