

**Article:**

Yifei K Yang.

Mining and Refining Elusive Microproteins from Noncanonical Open Reading Frames.
Clin Chem 2026; 72(10): 1101–3. <https://doi.org/10.1093/clinchem/hvag107>**Guest:** Dr. Yifei Yang is a clinical chemist and clinical lab consultant in the Department of Laboratory Medicine and Pathology at Mayo Clinic in Rochester, Minnesota.

Bob Barrett:

This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I'm Bob Barrett.

The traditional understanding of the human genome states that it encodes roughly 20,000 protein-coding genes, with further complexity attributed to alternate splicing and post-translational modifications. Recent work using Ribo-Seq and immunopectidomics provides compelling evidence for the presence of an entirely new class of genes, those that encode microproteins. As the name implies, microproteins are much smaller than normal proteins and seem to have originated much more recently in the process of human evolution. Although there is still much to learn about this new class of proteins, we do know that they are involved in a whole host of biological processes, from regulation of cellular metabolism to immune surveillance. A News & Views article in the October 2026 issue of *Clinical Chemistry* highlights recent developments in the discovery and characterization of microproteins.

How have technological advances made microprotein discovery possible? How do regular proteins and microproteins work together to make our bodies function? And will these microproteins become associated with a disease state, ultimately transitioning to analytes we routinely measure in the clinical lab?

Today, we're joined by the article's author. Dr. Yifei Yang is a clinical chemist and clinical lab consultant in the Department of Laboratory Medicine and Pathology at Mayo Clinic in Rochester, Minnesota. She is interested in expanding our understanding of immune competency and regulation through diagnostic biomarkers. So, Dr. Yang, in the News & Views article, you reported on a recent *Nature* publication on expanding the human proteome by annotating microproteins. So basically, what are microproteins in the context of human biology of genome and proteome?

Yifei Yang:

So, I appreciate the opportunity to expand on this rather complex paper, which I may not fully cover in this mini-review. So, for listeners and readers who are not in the

genomic field, maybe we want to start with the Human Genome Project, which is this big project I believe started in the 90s that wants to sequence the human genome. And then they essentially primarily sequence the protein coding genes. In this case, they are essentially verified canonical open reading frames. And these open reading frames, or canonical open reading frames, are long sequences, DNA sequence, with exons and introns, with the start codon, the stop codon, and your different upstream and downstream regulatory elements. So these are the canonical genes, or the human genome, that we now know and utilize every day for disease detection, diagnoses, treatment, et cetera.

And then if we follow the central dogma, you want to first define the DNA sequences, and then you want to find their corresponding RNA transcripts. And these RNA transcripts will then be translated into protein sequences or protein products. And they are what we consider to be the canonical proteins in the context of this specific publications. Hence, you have the human proteome, which is essentially the collection of all proteins expressed and translated by human cells. And then I guess one thing to note is that DNA to RNA transcript is not a one-to-one connection. You can have multiple RNA transcripts from one protein coding genes. And then from RNA transcripts to protein is also not a one-to-one correlation because you can have all these post-translation modifications of proteins that are not tied to your DNA sequence.

So that's why you most likely have more collections of proteins in your human proteome compared to your human genomes. But I guess that is true with modern science is that, or past science too, is that anytime that you think that you have mapped out the entire proteome or genome, there is like the dark side, or the side that you haven't discovered. So we have the dark genome, which are essentially these difficult to sequence region that are not extensively sequenced compared to the protein coding sequence of the DNA. And we also have essentially this, I don't know if it's necessarily called a dark side or a gray side of the proteomes, which are the microproteins that we're going to discuss today. And essentially these are not canonical proteins and their evidence of existence is actually first uncovered at the RNA level.

I guess, again, with the technology advancement when it comes to like molecular sciences, we now can actually sequence RNA sequences that are specifically associated with ribosome. Again, ribosome is responsible in converting or translating this messenger RNA sequences into protein sequences. So we think these ribosome-associated RNAs are actually encoding for proteins, but they are not the traditional messenger RNA that we have seen and profiled. And in

addition to these ribosome-associated RNAs, I guess in this particularly review and also the *Nature* paper, they are called Ribo-Seq RNA sequences. Researchers have also uncovered growing numbers of microproteins, which they think are the product of these non-canonical Ribo-Seq RNA sequences. So that's what we consider to be the microproteins. And essentially these are proteins that have not been profiled and have not been studied, extensively studied before.

Hence the hypothesis is that we want to annotate the function of these microproteins. And we also want to tie or link these microproteins to the ribosome-associated non-canonical open reading frames. But I guess the question still remains about like how to link the two molecules together, because if you follow the central dogma, then you need the DNA, RNA, and the protein all at the same time. In this case, the DNA sequence will still be the same. And then we have the Ribo-Seq RNAs, and then we have the microproteins. So this is essentially what this *Nature* paper is about, is trying to connect the three together.

Bob Barrett: Okay, well, let's talk about those microproteins and how they are different from normal proteins, or as you call them in the paper in here, canonical proteins. Why are they so elusive to detect and identify?

Yifei Yang: Okay. So first, microproteins, I guess you probably can refer from the name that they're relatively short polypeptide. Again, this is in contrast to the canonical proteins or the canonical genes, because they are normally considered to be these DNA sequences with thousands of base pair, and their related protein product will be more than 100. I mean, definitely we have smaller canonical proteins as well. They're normally considered to be long peptide that can form their own three-dimensional structure and has functions. In comparison, microproteins are relatively short. I think they're normally the one that they've identified so far are shorter than 50 amino acid long.

And another very important aspects of them is that, again, first of all, their sequence are very different from the canonical protein. So, we know that they're not just a degradation or a cleavage product of canonical proteins. And also they're less evolutionary conserved from both protein sequence and a DNA sequence perspective, meaning most of the functional canonical proteins that we know today, they are conserved across a different species, meaning we may share, I'm just making up a number, but let's say like 90% of the sequence similarity with other coordinates or other mammals. I guess we've all heard that we're only a couple of cells in base pair differences from apes. But anyways, from the perspective of microproteins, they are considered to be less evolutionary conserved. Hence, you may not see that

high level of similarities across different species when it comes to their DNA sequences.

And I don't believe that we've extensively studied microproteins from other species. So then we really can't compare from a protein level across species. The reason they're also elusive to detect and study is because they exist in our cells at a much lower abundance. And I think this is still like the hypothesis or the assumption, but I think they're believed to be expressed and translated transiently in nature. So they don't function as a constitutional protein, like the protein we have, like the protein in our bones or the proteins in our skins, that they exist at all time. But I think these microproteins are relatively low abundance and transient in nature, hence they're difficult to detect and study.

Bob Barrett: So what biological roles do they play and why are you interested in learning about how they're annotated?

Yifei Yang: What interests me is actually, particularly because their elusive nature and it is inherently difficult to fully determine a function of a microprotein. So I am a clinical chemist. So we study and measure biomarker that has been validated and has shown significance in a disease diagnosis or prognosis. So for us, it's always an interesting perspective to look at this biomarker that are still early in their discovery and in understanding how viable some of these results that you see early in discovery will turn out on the clinical aspects. And I guess maybe this goes a little bit more on how a protein function can be annotated.

So similarly, you wanted evidence to be at DNA, RNA, and protein level. When you determine a protein, you can knock out that protein from a DNA level or knock down the protein from the mRNA level, and then you can study what are the consequences of not having this protein in the cell or in a bigger organ system or in an animal model. But I guess it is inherently difficult to do these kinds of studies in microproteins just because we first don't always know what are their corresponding DNA or RNA molecules and how we can essentially knock down or knock out the specific gene and specific expression.

And if you read about the field, there appears to be actually quite a bit of controversy on what constitutes a true functional annotation of a microprotein. But I guess if you read the studies, they're mostly what we call like a correlation study. So you see an upregulation of a specific microprotein or their associated RNA transcript and what are the consequences from like a cellular level or maybe from a disease model perspective. So they're what we consider to be correlation evidence. So they're not a causality evidence, but they appear to be associated with RNA transcription and their

process. And these correlation studies have shown that they are involved in cell metabolism, immune surveillance, and cancer development. So there's still great interest from a biological or disease perspective to study these molecules. But I think the field may have to think beyond what we consider to be a canonical functional annotation to study these microproteins.

Bob Barrett: Doctor, how are immunopeptides different from other signature peptides used for protein identification?

Yifei Yang: Yes. So this touched on an important level of evidence that we have to detect and identify an existence of a protein, or in this case, a microprotein. So normally to detect a protein, you can visualize it based on antibody but in this particularly *Nature* paper, they're primarily dependent on a mass spectrometry-based evidence. So to study a protein using mass spec, you can first digest a protein based on proteases, and these proteases will cut the protein sequence at a specific location. So then you generate a collection of peptides, and you can find a signature peptide that are specific to this protein. So if you detect the signature protein, or I guess by definition of the Human Proteome Project, you actually have to detect at least two of these peptides, and then you can say, "Oh, I detect the presence or the expression of a specific protein."

And on the other hand, we talk about because of the microproteins are expressed at low level, it may be difficult to detect microprotein when they only exist in like a nanomolar concentration when your canonical proteins are expressed at millimolar or micromolar concentrations, right? So then the researchers in this *Nature* paper looked at immunopeptide. Immunopeptide is actually generated by our own immune system, specifically by our own antigen-presenting cells. So the antigen-presenting cells will look at proteins, and then they will engulf it, and then they will process it with their own intracellular organelles. And then they produce these immunopeptides, and then they present these immunopeptides either as our own protein or extraneous protein on these MHC [major histocompatibility complex] molecules on the cell surface. So then immunopeptide inherently, you can enrich for immunopeptide by enrich for these MHC molecules, so then they don't have to overcome these huge concentration differences compared to the canonical proteins.

So then you can actually find these microproteins or their related peptides more easily by looking at the collection of immunopeptide experiments. But I guess the shortcut is that they are processed differently by like different antigen-presenting cells, and then they may be different in different people as well. So they are less specific compared to a

proteases-generated peptide. So then you won't be able to say, "Oh, specifically, this is a signature peptide for this protein." So from a Human Proteome Project perspective, they view the immunopeptide as a weaker evidence, or weaker level of evidence, compared to the signature peptide that you generated from these trypsin proteases digestion experiments.

Bob Barrett: Well, finally, Dr. Yang, as a clinical chemist, are there any clinical implications for identifying and understanding these microproteins?

Yifei Yang: So, as I had mentioned earlier that on the clinical side, it's very interesting to look at these discovery-based biomarkers, a new biomarker that is implicated in a disease pathogenesis or maybe potentially a treatment target. But based on past experience, we can see that it normally takes a lot more clinical validation studies to translate these discovery-based biomarkers into clinical practice. I mean, the timeline could really vary, but I would say at least like five years or maybe even a decade. And on top of that, we need very sensitive and robust methods to detect and quantify these biomarkers if they're going to be used for clinical management.

And because of these microproteins are so elusive and their biological or pathological function have yet to be fully established, I see that there may be evidence to become a clinical biomarker is still emerging. But it does seem like these microproteins, again, because they're strongly associated with these RNA regulatory motifs, so they may be very important target or could be future targets as a therapeutic target if you want to target a certain gene, you may want to target a microprotein that may be associated with this gene's regulation. So, they could be a very effective therapeutic target in gene therapies or canonical or non-canonical gene therapies.

And if they do become a therapeutic target themselves, then we may need diagnostic assays to ensure that these targets are being hit as intended, and what are the efficacies of these targets are being utilized. So, I will be keeping a close eye on the field and see how new studies come out or change our view of our current human proteome and human genome as well.

Bob Barrett: That was Dr. Yifei Yang from the Mayo Clinic in Rochester, Minnesota. She wrote a News & Views article in the October 2026 issue of *Clinical Chemistry* highlighting the characterization of microproteins, and she's been our guest in this podcast on that topic. I'm Bob Barrett. Thanks for listening.