

Thousands of new proteins revealed in dark proteome

Scientists have uncovered more than 1,700 new proteins that could have implications for human diseases, including cancer. Mostly very small, these proteins were found in what's called the 'dark proteome', which covers gene products from previously overlooked sections of DNA. These proteins have unusual properties, motivating scientists to coin a new concept, peptideins, to help understand their potentially unique biology. Their findings are being shared with scientists worldwide in an open-source format to stimulate further research.

Genes in DNA provide the recipe for cells to produce strings of amino acids, called peptides. Historically, peptides have been called proteins if they are long enough and have existing evidence for a biological role, such as the appearance of the same protein across species in evolution. A large, curated international database of proteins contains some 19,500 entities. But increasingly, scientists believe the traditional definition of a protein needs to be broadened.

A team of scientists led by the Princess Máxima Center for pediatric oncology, the University of Michigan Medical School, the EMBL European Bioinformatics Institute and the Institute for Systems Biology looked at more than 7,200 previously understudied sections of the DNA called non-canonical open reading frames (ncORFs). They found that some 25 per cent of these sections – more than 1,700 – generated detectable protein-like molecules. These proteins are smaller than traditional proteins and therefore referred to as 'microproteins'.

The new study was published in the prestigious journal *Nature* today (Wednesday). It was supported by funders including the US National Institutes of Health, the National Science Foundation, Onco Accelerator (a Dutch National Growth Fund program) and the European Union's Marie-Sklodowska-Curie program.

In the new study, scientists looked at 3.7 billion individual bits of raw data that may support known and previously unknown proteins – drawing upon 95,520 experiments, which took around 20,000 hours for computers to complete working non-stop. They found 1,785 microproteins, a number that at first glance would increase the protein databases by nearly 10 per cent.

But most of these 1,785 microproteins didn't resemble the other 19,500 traditional proteins. For example, they were very small: 65 per cent were fewer than 50 amino acids in length, compared to less than 1 per cent of the 19,500.

Looking more closely at the microproteins, they saw that only a few – perhaps a dozen – resembled the traditional proteins. For the remaining bulk, they spent over a year trying to figure out how to make sense of them.

Working with protein experts from across the globe in the TransCODE consortium, the scientists coined a new biological concept: the peptidein. For decades, the research community has had a binary view of the relationship between human DNA and human proteins. A given piece of DNA either *does* or *does not* produce a protein. In their new study, the scientists propose a third choice: DNA could make a protein, a peptidein, or neither.

The team defined a peptidein as existing in cells as a protein-like molecule, meaning that it is made of amino acids like proteins are. But the role of a peptidein is ambiguous. Perhaps it has a function in normal human biology, or perhaps not; this is the key distinction with

traditional proteins, where all are believed to have a function in normal human biology even if the details of that function are not fully known yet.

Importantly, this definition of peptidein leaves the door open for it to become a 'protein' in the future – that is, if scientists gather more evidence on it. To start exploring this idea, the team searched for peptideins without which cells cannot survive. These so-called pan-essential peptideins can be important candidate drug targets in cancer and other diseases.

Using large-scale CRISPR gene editing, the scientists found six peptideins that looked promising. For example, one of these was a peptidein produced from OLMALINC, a genetic sequence previously thought not to produce proteins. When the researchers switched this gene off, 85 per cent of more than 485 cancer cell lines showed impaired survival. The researchers confirmed that this effect comes from the peptidein itself, not the RNA molecule it sits on, and found that it plays a role in cell division and DNA damage response.

Many of the newly detected peptideins are presented on cell surfaces for recognition by the immune system, making them potential targets for cancer immunotherapy. A number of such molecules presented to the immune system are already under development as drug targets, and there is growing interest from both academia and industry in exploiting this new class of cancer antigens. Peptideins could also shed light on genetic diseases that conventional gene analysis has been unable to explain, simply because genetic diagnostics were unaware that these molecules were encoded by the human genome.

Members of the consortium had previously uncovered an essential role for a microprotein, ASNSD1-uORF, in children with a high-risk form of the brain cancer, medulloblastoma. Scientists at the Princess Máxima Center are now carrying out further research to determine its role in additional pediatric cancers with the activated MYC oncogene, such as neuroblastoma.

Dr. Sebastiaan van Heesch, research group leader at the Princess Máxima Center for pediatric oncology, who co-led the study, says:

"We know that the current overview of recognized proteins doesn't capture the full picture. With this study, we show that thousands of overlooked genetic sequences contribute to the dark proteome by producing a new class of protein-like molecules, microproteins, that had been missed before now. But for most of them, we don't yet know what they do.

"It felt really special to discuss and decide what to do with this new class of molecules, as we had gathered enough early evidence to suspect that they might be widespread across cell types and tissues. By classifying these molecules of unknown functionality as peptideins, we've given them a formal place in reference databases so the wider community can study them.

"With growing interest in industry and academia, peptideins are at the center of multiple drug development initiatives. Similarly, we see them increasingly turning up as important players in diseases, including childhood cancers. We hope to inspire a new wave of research into peptideins and to unlock new insights and drug targets across human biology, particularly for the development of cellular immunotherapies and cancer vaccines."

Dr. John Prensner, pediatric neurooncologist at the University of Michigan Medical School, who co-led the study, says:

"We're just beginning to see what this 'dark proteome' has to offer. It's like the trailer to a movie. We see the outline of a game-changing view of human biology. We're incredibly

excited that the coming years will open new doors to help solve and treat human diseases such as cancer.”

Dr. Robert Moritz, Professor and Head of Proteomics at the Institute for Systems Biology, who co-led the study, says:

“Our collaborative work represents a culmination of decades of investment from federal funding agencies in building the computational and data infrastructure needed to interrogate the proteome at truly unprecedented scale at the Institute for Systems Biology. By deploying our battle-hardened Trans Proteomic Pipeline across nearly 100,000 mass spectrometry experiments encompassing 3.7 billion spectra — derived from the world's collective publicly available mass spectrometry data, with the results housed within PeptideAtlas at ISB for the scientific community to view and share — we were able to confirm, with high confidence, the existence of more than 1,700 of these newly identified peptideins that would otherwise have largely remained invisible to science.

“What excites me most is not simply that these molecules exist, but what their existence implies.

“Biology has long relied on a relatively small cast of well-characterized proteins to explain the regulatory logic of the cell, but peptideins suggest that beneath that familiar layer lies an entire untapped layer of molecular actors whose functional roles in gene regulation, signalling, and cytopersistence, many we are only beginning to imagine. Given their smaller size and the diversity of cellular contexts in which they appear, I believe peptideins may prove to be among the most versatile and consequential regulatory molecules we have yet encountered in human biology. This is not the end of a search — it is the opening of a vast and fertile new territory for the entire scientific community to explore and exploit, and I look forward to seeing what the broader scientific community uncovers as these molecules, and many more that are yet to be confirmed, are brought into the light.”

ENDS

Notes to editors

For more information about the study or to request an interview with one of the scientists, please contact the Princess Máxima Center for pediatric oncology's press office.

Press manager Elco van Groningen: e.c.vangroningen-2@prinsesmaximacentrum.nl or +31650173714

Science communications officer Sarah Wells: s.wells@prinsesmaximacentrum.nl or +31650006607

About the TransCODE Consortium

The study is the work of the TransCODE Consortium, an international collaboration of more than 60 researchers at over 30 institutions worldwide, co-led by the Princess Máxima Center for Pediatric Oncology in the Netherlands, the University of Michigan Medical School, the EMBL European Bioinformatics Institute in Hinxton, and the Institute for Systems Biology in Seattle.

About the Princess Máxima Center for pediatric oncology

When a child is seriously ill with cancer, only one thing comes first: a cure.

That is why at the Princess Máxima Center for pediatric oncology, we work together every day in a groundbreaking and passionate way to improve the survival rate and quality of life of

children with cancer. Now, and in the longer term. Because children still have a whole life ahead of them.

The Princess Máxima Center is a research hospital, the largest pediatric cancer center in Europe. All children with cancer in the Netherlands are treated here. Our mission: To cure every child with cancer, with optimal quality of life. Over 450 researchers and 900 healthcare professionals work closely with Dutch and international hospitals on better treatments and new perspectives on cures.

In the Princess Máxima Center, Utrecht, Netherlands, the best healthcare professionals and scientists work together on a unique mission: to cure every child with cancer, with optimal quality of life. A mission that remains urgent, as one in four children with cancer dies, and many children who do survive experience the effects of radical treatments throughout their lives.

About the University of Michigan

One of the nation's top public universities, the University of Michigan has been a leader in research, learning and teaching for more than 200 years. With one of the highest research volumes of any public university in the country, U-M is advancing new solutions and knowledge in areas ranging from the COVID-19 pandemic to driverless vehicle technology, social justice and carbon neutrality. Its main campus in Ann Arbor comprises 19 schools and colleges; there are also regional campuses in Dearborn and Flint, over 200 centers and institutes, including the Center for RNA Biomedicine, and a nationally ranked health system, Michigan Medicine.

About ISB

The Institute for Systems Biology (ISB) is a collaborative, cross-disciplinary, non-profit biomedical research organization based in Seattle. We focus on some of the most pressing issues in human health, including aging, brain health, cancer, chronic illness, infectious disease, and more. Our science is translational, and we champion sound scientific research that results in real-world clinical impacts. ISB is an affiliate of Providence, one of the largest not-for-profit healthcare systems in the United States. Follow us online at isbscience.org, and on YouTube, Facebook, LinkedIn, X, Bluesky, and Instagram.

About the EMBL European Bioinformatics Institute