GenomeWeb/Mendelspod PODCAST SERIES

genomeweb

mendelspod.com

Podcast host Theral Timpson and Illumina Distinguished Scientist Kyle Farh. Photos courtesy of their subjects

Podcast: Genomic AI millions of years in the making

Illumina VP and Distinguished Scientist Kyle Farh spoke with Mendelspod *about using natural selection to train gene-identifying algorithms*

THE PODCAST MENDELSPOD covers the latest developments in biotech and precision medicine. For 13 seasons and more than 10 years, host Theral Timpson has conducted longform interviews with scientists, executives, and journalists about the biggest ideas and newest technology driving these fields.

Illumina is proud to share that, in partnership with GenomeWeb, our Vice President and Distinguished Scientist Dr. Kyle Kai-How Farh appeared as Timpson's guest on the most recent episode.

Listen to the Mendelspod *episode "How Do You Train Genomics AI? On Natural Selection Itself, Says VP of Illumina's AI Lab Kyle Farh" here: https://gw-resources.genomeweb.com/free/w_illu09/prgm.cgi?a=1*

Farh earned his medical degree from Harvard Medical School and his PhD from MIT, and his dedication to genetic science has been the throughline of his career, from postdoctoral medical and population genetics studies at the Broad Institute to his residency in the Clinical Genetics department of Boston Children's Hospital.

For the past seven years, Farh has led Illumina's Artificial Intelligence Laboratory for Genome Interpretation, and he has contributed to 10 articles in *Cell*, *Science*, *Nature Genetics*, and other prestigious journals—the most recent of which covered his lab's development of PrimateAI-3D, an algorithm that's effectively trained by millions of years of natural selection to identify potentially pathogenic gene variants.[1]

Studying our relatives to discover ourselves

In the podcast, Farh discusses genomic AI in detail. Whereas the most well-known AI programs can draw from a wealth of published works to make their predictions (in ChatGPT's case, predicting which word is most likely to come next in a sentence), AI built to identify previously unknown effects of gene variants has no such luxury.

"If you say, 'In the human genome—base 2,000,000,006—A turned to C. Does that cause disease or not?' There's no human who can actually really tell you that," Farh explains. "There's vast amounts of data,

1. https://www.illumina.com/company/news-center/feature-articles/PrimateAI-3D.html

but for the most part, there [are] no labels for what the correct answer should be."

So to make the best predictions for human health, his team instead sequenced over 800 individuals from 233 species of nonhuman primates. They used this data to train their algorithm to identify variants that are common across our closest relatives in the tree of life—if those variants have survived millions of years of natural selection, it's safe to rule them out as benign. And by identifying benign variants, the process of elimination makes it easier to find potentially pathogenic ones.

Farh notes that ClinVar, a widely used public database, previously covered about 70,000 gene variants—this method of machine learning expands the number of variants interpreted to almost 4.5 million.

Breakthroughs don't come easy

Farh credits their ability to conduct this research to many factors unique to Illumina: The company had existing relationships with over 70 academic authors around the world ("Most of them are people who actually went out to the jungle and caught the monkeys to provide the samples"). Illumina has the technology to generate vast quantities of sequencing data from those samples, and it has funding for the computing power necessary to interpret that data. "Deep-learning compute has really become very unaffordable for many labs in academia, at least at the scale needed to get the most powerful, most effective models," he says.

One of PrimateAI-3D's other breakthroughs is in its name: Farh's team can use it to superimpose newly identified benign variants from the DNA sequence onto the three-dimensional protein structures they code for. "From that, we can figure out which are the pockets where the pathogenic variants are," he explains. "A lot of times, these pockets of parts of the protein are very obvious in 3D space, whereas they're quite disconnected in linear space in the genome."

The Illumina AI lab has already demonstrated strong correlations between PrimateAI-3D's predictions and real-world effects. "You can show that someone's blood cholesterol level is very well predicted by their PrimateAI-3D score for the variant that they carry in the *LDL* gene, or in the *PCSK9* gene," he says. "This allows you to quite accurately predict who will be at risk for diseases like dyslipidemia or type 2 diabetes," as well as for less common variants.

PrimateAI-3D is now or will soon be available across Illumina's software products—Farh specifically mentions DRAGEN Secondary Analysis, Illumina Connected Analytics, and Emedgene variant interpretation software for genetic disease applications.

Predictions of a more speculative nature

When asked about what's next for his laboratory, Farh offers a few promising leads: His team is experimenting with "perturb-seq," a gene-editing technique that introduces different mutations to individual cells and then measures how those mutations affect cell function. ("That's super fun," he notes.)

They're also working on algorithms to identify pathogenic variants in the parts of the genome that don't code for proteins. Timpson suggests that this could move the needle on rare disease diagnosis, and Farh agrees: "Right now, the truth is, our diagnosis rate with exome [sequencing] is 30%. And that's very unsatisfying, I think, for patients and physicians."

Over the next five to 10 years, Farh sees a new era of precision medicine coming, facilitated by efforts currently underway to sequence cohorts of unprecedented scale—he points to the relatively recent discovery that "people who carry natural mutations that break the *PCSK9* gene have naturally very low levels of cholesterol and are protected from heart attacks.[2] So very quickly, pharma was able to make antibodies which inhibit *PCSK9*—and now that's a fantastic drug. So I think that there [are] many, many other genes like this out there where natural-occurring mutations are beneficial and are great potential therapeutic targets for all kinds of diseases. And ultimately, AI is actually also central for identifying those." ◆

To read more about how PrimateAI-3D powers genomic variant annotation in Illumina Connected Annotations, read this article on Illumina's Genomics Research Hub: https://www.illumina.com/science/genomics-research/articles/primateai-3d.html

2. https://www.nature.com/articles/496152a