Welcome to *MIT's Computer Science and Artificial Intelligence Labs Alliance's* podcast series. My name is Steve Lewis. I am the assistant director of Global Strategic Alliances for CSAIL at MIT. In this podcast series, I will interview principal researchers at CSAIL to discover what they're working on and how it will impact society.

Manolis Kellis is an associate professor of computer science at MIT. A member of the Computer Science and Artificial Intelligence Lab and the Broad Institute of MIT in Harvard, where he directs the MIT computational biology group. He has received the US Presidential Early Career Award in science and engineering for his NIH R01 work in computational genomics. He was recognized for his research in genomics is one of the top young innovators on the age of 35 by *Technology Review* magazine, and one of the principal investigators of the future by *Genome Technology* magazine.

He received the Gregor Mendel Medal for outstanding achievements in science by the Mendel lectures committee, and was also listed as one of three young scientists representing the next generation in biotechnology by the Boston Museum of Science. He obtained his PhD from MIT, where he received the Sprowls Award for the best doctorate thesis in computer science and was the recipient of the first Paris Kanellakis Graduate Fellowship.

Prior to computational biology, he worked on artificial intelligence, sketching image recognition, robotics, and computational geometry at MIT and at the Xerox Palo Alto Research Center. Since you last joined us back in February of 2020, a lot has happened in the world, especially pertaining to public health. How has this affected your research into disease and public health?

So there's really been just a dramatic shift in the whole world. So basically, we saw our world change in front of our eyes and it has affected so many people, so many lost loved ones, so many lives lost, so much livelihood lost. At the same time, we saw the best of humanity come together. We saw scientists across industry, academia, health care really coming together as a joint force to do everything we can to address this pandemic.

So every person putting their best technologies, their best ideas to mind, to address as a common humanity this plague that is killing so many of our loved ones. And in the world of genomics in particular, there was this dramatic speed up in the sharing of information and the dissemination of information which was truly a glimmer of hope amid so much tragedy. So if you look at the speed with which the vaccine were developed, the genome was first published in late January. And the first vaccine design was completed less than 40 days later.

And the first production started happening on the order of months. And the approval was done in less than 9 months from the design of the vaccine. This is unprecedented. Vaccine development typically requires years and sometimes decades to shall go through all of the layers of designing the vaccine learning about the new pathogen identifying the pathogen itself figuring out its genome, et cetera. So we've seen the power of genomics really come to light with this pandemic.

Our own group has an expertise in genomics in understanding genomes in understanding what's inside genomes. And we also just like every scientist in the world try to contribute to the best we can to the understanding of this pandemic. So the first thing we did immediately last year in February and March is use our comparative genomics techniques for understanding the SARS-CoV-2 genome, the genome of the pathogen of the virus that underlies the COVID-19 pandemic.

So you would think that a genome so small of only 29,000 nucleotides would be trivial to understand and would already be well sorted out. The human genome that our group has worked to annotate is actually 3 billion nucleotides long. So we're talking about four orders of magnitude difference between the two, actually five orders of magnitude. So it's an enormous challenge to basically anotate the human genome. And we've developed some extremely sensitive techniques for using comparative genomics and using something that we like to call evolutionary signatures for understanding where are the protein coding segments of the human genome.

So the signatures are basically looking at how nucleotides and how specific codons, which are triplets of nucleotides that encode a single amino acid. How they evolve across related species? And what we use there is the fact that the evolutionary constraint is acting at the level of protein. And therefore, the underlying nucleotides are free to change as long as they preserve that selected function.

In the case of a protein coding region that gives you a very specific signature of protein coding like evolution that we are able to look at across the 3 billion nucleotides of the human genome to discover hundreds of new human genes, hundreds of new human axons to reject previously hypothesized genes that are clearly not protein coding. And we took that expertise and we applied it to the SARS-CoV-2 genome that had just recently published.

So we basically went out to the public databases and we identified another 44 closely related species or strains if you wish. So species what we use for organisms like us like yeast, et cetera, but we use strains for viruses. So we basically looked at multiple other RNA genomes that are closely related. And when I say closely, I mean, at roughly the distance of human to other mammals. So this is roughly 60 million years of divergence in human, but in viruses, it's obviously is much more condensed because viruses move much, much faster.

So what we found there is that we could actually have a very clear signal that tells us what are protein coding regions in the SARS-CoV-2 genomes and what are not. And the results were quite striking. We could see the beautiful conservation patterns of the first 16 protein coding genes that are all encoded in a giant open reading frame and that are spliced from each other post-translationally. We could see the spike protein evolving super, super rapidly and yet showing very clear signals of protein coding evolution.

But we also found some surprises ORF-10, which is the last gene of the genome that has been annotated as protein coding for many, many related species, related strains of coronavirus turns out not to have any protein coding constraint. So we were able to show that in fact, one of the quote unquote, "protein coding genes" that people had annotated among the 29 protein coding genes. It's not like there's thousands like there's a human. There's only 29% protein coding genes turns out one of them, which is bogus. It's not a protein coding gene at all.

So we now are able to look at RNA elements, functional elements within this region that are clearly functional or clearly selected evolutionarily, but at a nucleotide level not at the protein coding level. We had another surprise that ORF-6 and ORF-8, which are known to evolve very rapidly were in fact barely conserved at the nucleotide level. If you looked at the nucleotide evolution in other programs like phyloFi, for example. You would see that, there's really no signal there for protein coding constraint, but in fact, using our-- for nucleotide level constraint, but in fact, using our method, we saw that there was very clear signature of protein coding constraint.

So this is the converse you basically have. Almost no nucleotide conservation, but very strong protein coding conservation. Then another very big surprise came from two overlapping reading frames. So that's basically when encoded within the nucleotide sequence of one protein coding gene, you can almost write in the margins or write between the lines by shifting their reading frame of translation by off by one. So proteins are translated so genes are translated every three nucleotides into proteins every three nucleotides is one codon.

And therefore, in the same transcript, you can have one protein coding gene in coding. And you can have a different protein encoded by shifting by one and a yet a third one by shifting by two. And viruses do that a lot because they have very compact genomes. What's different about SARS-CoV-2 and many of these servico viruses, these are related beta coronaviruses. And coronaviruses more generally is that they have much larger genomes. So they have better proofreading, and that gives them the ability to have better larger genomes and bigger envelopes and stuff like that.

So that basically allows the genome to be larger. So they tend not to have as many overlapping genes, but in this particular case, overlapping the nucleocapsid protein that the RNA genome wraps around and is held together and that also plays roles in immune evasion and so forth. So overlapping the end protein which is also known as ORF-9. There's another ORF, ORF-9b that also showed very clear signal of overlapping constraints with two different reading frames were considered at the same time. So ORF-9b was previously hypothesized.

And in fact, we confirm that indeed there's an overlapping signal there. And there were other overlapping ORF that had been previously being hypothesized. And we found that none of them in fact shows protein coding constraints suggesting that only 9b from the previously hypothesized open reading frames is actually protein coding. And the big, big, surprise came within ORF-3a.

So ORF-1 is this giant open reading frames that encode 16 proteins or 11 proteins depending on an internal frame shift. ORF-2 is the spike protein. ORF-3a is in the next protein after that. And what we found is what we call ORF-3c, which is a new never previously seen protein coding gene that is hiding within ORF-3a. So what we basically are able to do is look at this genome and just use our other techniques to revisit it, to revise it, and to now give the community a new basic annotation of the SARS-CoV-2 genome that allows us to now go and study systematically these novel proteins in ORF-3c, for example.

The overlapping protein of ORF-9b to focus on the RNA functions overlapping ORF-10, which does not encode a protein and use again all that in fully public knowledge. So this is something that we posted on bio archive immediately as soon as we found the result. And this was eventually published a year later in Nature Communications, but this is something that, again, the whole community has been posting these publicly in a giant sort of collaborative family.

The other thing that happened through this effort is that we notice that different papers we're using different names from any of these overlapping genes. For example, ORF-3b and ORF-3d were confused in many papers. And some people were referring to one when, in fact, meaning the other and so on and so forth. So we actually brought the whole community of genomics lists for coronavirus genomes. And we wrote another paper that sort of proposes a reference naming for all of the genes to sort of resolve all these ambiguities and all these errors.

And again, that was just beautiful to sort of see how responsive everyone was in the community and how generously everyone was giving their time to come together and write this revision. The next thing we did is actually exploit the huge number of variants that have been isolated. And the mutations that have been called on these variants of the different isolates of the current pandemic to study how is the genome evolving today right now, compared to how it has been evolving over these millions of mammalian equivalent evolution?

And what we found was quite striking. The speed with which different genes are evolving in evolutionary time is quite predictive of the speed with which things are evolving today in the current pandemic. So genes that are just really fast evolving in general are fast evolving in the current pandemic. And genes that are very slow evolving are slowly evolving in the current pandemic with two exceptions. One exception is actually the spike protein that everybody has been, of course, talking about.

So spike S1, which is the first part of the protein that sort of binds the ACE2 receptor and then attaches to the whole cell before the S2 part of the protein basically enables insertion of the RNA genome inside our cells. What we found is that the S1 portion of the protein that was very, very fast evolving between different strains between different-- distantly related genomes, closely related genomes is in fact, much slower evolving in the current pandemic.

And this could simply mean that the adaptation of this one to the human host is quite a good match. And that because this pandemic is so quote unquote, "young" in the human species that we haven't yet started evolving our ACE2 receptor away from it. There's no natural adaptation to fight this virus as there is, for example, in bats. So between different bat species, there's this core evolution of these viruses that are hosted by the bats.

And there's millions of years of co-evolution between the two. So there's a lot of opportunity for adaptation between the host and the virus, whereas in this particular case, I mean, in human timescale, a year and a half is nothing evolutionarily. So basically, we're a very young species for this. And our immune system hasn't built defenses against the spike protein and other proteins yet. So the virus basically, early in the pandemic is able to freely change without having to and adapt all of each other proteins without having to adapt this S1 protein that appears to be very well adapted already to the human host.

So that was the first surprise. The second surprise was the nucleocapsid protein, which was super slowly evolving in different coronaviruses is in fact, much faster evolving in the current pandemic. And in fact, the fastest evolving region of the genome is exactly a region of the nucleocapsid protein which is overlapping an epitope for our immune system. So that might suggest that nucleocapsid is in fact evolving rapidly to avoid detection by the human host.

[INTERPOSING VOICES]

I was going to ask you, I like to just touch on sort of the variants that are in the mutations that are evolving with this virus. How has your research sort of helped identify that or maybe how can we get ahead of it? Is that something your team is looking at?

It's a great question. So this is again a story of collaboration. So we're collaborating with Nevan Krogan right now over at UCSF. What we're finding there is that if you start looking at the Alpha variant and the Delta variant, that there's, of course, mutations in the spike protein, but we're also seeing mutations in other proteins, including the new genes that we're predicting. The new protein coding genes that were discovered. And what we're seeing there is, for example, if you look at one of the ORFs that I mentioned earlier namely ORF-9b versus the nucleocapsid ORF.

What Nevan Krogan group found is that there's an increase in the translation of 9b. There's an increased abundance of the 9b protein that overlaps nucleocapsid in the Delta variant and also in the Alpha variant. And they had postulated that perhaps this is due to a change in the transcription of the corresponding transcript. So let me make a quick parenthesis here to basically say how incredibly ingenious this genome is.

So let me describe what's happening when the virus first enters your cells. So as I mentioned, the spike 1 protein attaches the spike 2, basically the second portion of the protein basically leads to conformational change that will then open up your membrane and insert a single messenger RNA from that virus. OK. We have 20,000 protein coding genes. We have hundreds of thousands of different transcript isoforms.

And in any one moment in our cells, there's really thousands upon thousands of RNA molecules floating around. We're talking about the insertion of one RNA molecule now. That's basically coming into a human cell, which is hundreds of times its size and is going to hijack and take over that cell. I mean, mission impossible has nothing on this genome in terms of one infiltration taking over an entire society, which is our cell. What happens? So the mRNA single innocuous mRNA enters the cell.

Our transcriptional machinery-- the virus has no transcription machinery. Our transcription machinery will see this positive stranded RNA molecule and say, all right, seems like it's my job to translate, so this looks good to me. I'll translate it. And it will start translating. And every single time, an mRNA is seen by the translation machinery by the ribosome. The ribosome will basically look for a start codon and then start translating making amino acids. And it will find a stop codon and it will stop translating.

So the first 11 genes, 11 proteins are translated from that one mRNA. By starting translation with the ATG and then continuing and continuing and continuing. And as this one protein gets formed, it starts folding onto itself inside our cells, single mRNA. And the third part, the third protein that is encoded in that very giant open reading frame which is like 16,000 nucleotides long. So one of the longest proteins to be translated inside ourselves. It will start creating these folds in these domains, and the third one is actually a cleavage protein that will basically go around and grab onto the other ones and cut them.

So you now have this cutting of multiple of 11 different protein products that's happening from this one mRNA translation. It's remarkable. So basically, there's like a Trojan horse. One person enters and then they free up all their partners and now you have 11 proteins in there. The second thing that happens as these proteins are now translated, they start changing our own cells. They start marking up all of our other mRNAs as, nah, this is no good, don't translate it.

So that the ribosome machinery of our cells is now focused almost exclusively in translating copies of the mRNA itself of the virus. So suddenly the cell is starting to be taken over. So the machinery is now redirected to making copies of that RNA. And while our own RNAs are starting to get degraded by our own machinery that basically says, oh, there's something wrong with this one. Let me degrade it. Because some of these proteins are going off and mark you with.

So that's the first 11 proteins. I mentioned that the first open reading frame can be either 11 or 16 proteins. How do the 16 happen? By having a programmed frame shift that causes the ribosome to skip by one nucleotide. And therefore, instead of meeting the stop codon ORF-11, it will basically continue and it'll now translate or 12, 13, 14, 15 and 16. So that basically creates another set of proteins now at a lower abundance. So ORF 1 through 11 is higher abundance than 12 through 16 is lower abundance. And all of these are now starting to do something very funky.

So remember how I told you that every protein, every RNA molecule will basically lead to effectively one translation. There's no internal reinitiation in the human genome. That's very rare. So our genes are monocistronic rather than polycistronic basically meaning that they only encode one protein at a time. So how do you translate ORF 2, 3, 4, 5, 6, 7, 8, 9, 10, well, not 10, but all the way to nine that are encoded after that giant opening frame.

The way that you do that, again, this is so beautiful. The genome itself contains these transcription regulatory sequences or TRSs intervening between right before ORF-2, right before ORF-3, right before ORF-4 and so on and so forth. And these will basically link up to the beginning of the normal transcript, the normal RNA of the virus. And they will cut off the intervening part so that they are now the first open reading frame of that transcript. And then same thing for the third one, he will become the first one by cutting off everything between before them.

So you now end up with these single mRNA that has now created all of these partner proteins. One of these partner proteins is basically grabbing the RNA cells, which is positive strand copying it onto the negative strand and then copying it again onto the positive strand and copying to get in on the negative and the positive. So if you end up with both positive and negative strand RNA, so it's basically making more copies of itself. And many of those are now getting truncated into what we call sub-genomic RNA, such that there's all of the surrounding all of the remaining open reading frames that are basically now going to make new proteins, including spike and so on and so forth. nucleocapsid.

So all of the proteins necessary for packaging the virus are basically ORF two through nine, whereas all of the proteins necessary for hijacking the cell are 1 through 11 and then 1 through 16. It will then go and make additional proteins that are now necessary for getting out of that cell and then taking over the rest of your body. So the first part is taking over that cell and now you're now constructing more copies of the virus. You're packaging it up with the nucleocapsid with the RNA wrapping around it. You're basically building the envelope.

And that's the envelope protein there's. The membrane protein, the E and M and then there's spike protein that basically creates this corona that we can look. So what we found is that the Alpha variant and the Delta variant, for example, are making much more of this ORF-9b. How's that possible? Again, this is the nucleocapsid transcript. The end transcript ORF-9a or ORF-9 simply. But every now and then, the signal that-- the ribosome reads to start translating here.

Every now and then that ribosome will miss the initial translation initiation signal. So every RNA has a start position for, of course, transcription. But then a start position for translation. And that ATG has a context what we like to call the Kozak sequence named after the scientists who discovered it. But the Kozak sequence of the translation determines with what efficiency the ribosome will start there. And if you have a lower efficiency sequence, if the Kozak score is lower, you will skip the first one and maybe start translating at the second amino acid.

So what's happening with ORF with variant Alpha and variant Delta is that they have what we discovered is that they have a lower affinity for translation initiation for nucleocapsid. And the nucleocapsid RNA is enormously highly transcribed. And that basically means that if you lower the affinity with which end gets translated, that means that 9b will get translated much more frequently. So there's a shift from end to 9b and there's a giant sort of increase in the translation of 9b. That is associated with both the Alpha variant and the Delta variant.

And one of the things that Nevan's group is trying to do now is figure out when does this do biologically? How does that go inside the cell? So this is one of the ways in which we are trying to understand the evolution of the virus. The first is looking at the mutations that are happening and interpreting these mutations biologically.

So for example, in the Nature Communications paper, we showed that the D614G mutation in the spike protein, which means that at position 614 of the spike protein, which is about 1,000 nucleotides long. There's a shift from a D amino acid to a G amino acid, which basically causes the strains that carry that mutation to increase in frequency. And we've seen that early in the pandemic. So basically, last year in March and April and May, we could see the increase of the D614G mutation happening in so many different countries independently. Wherever it was which is increasing in frequency.

And it became, of course, the dominant version in the pandemic. And basically, what we showed is that, in fact, that mutation had never before happened in the history of the 44 viruses that we had compared. And it was happening in the 11 amino acid stretch that had never seen a single change in any of the 44 genomes. Suggesting that perhaps this was an adaptation to the human host that was happening over and over and over again independently.

Basically, the two scientists number one understanding the mutations that are associated with these variants. And what do they do functionally based on our understanding of the language of DNA, the signals that the protein translation machinery that the transcription machinery are using to utilize this genome, that's number one. And then number two using our evolutionary comparisons in the lens of evolution to understand how these are changing across different in the dynamics of current day evolution of the virus.

Sort of shifting gears a little bit here, but related in the past you've studied obesity and cancer in our genomes. And since the outset of COVID-19, what are your thoughts about the interconnectedness of disease now with the world's focus shifted to comorbidities, specifically with obesity?

So we've studied obesity, we've studied Alzheimer's, we've studied, of course, SARS-CoV-2. And it's kind of crazy for a lab to be working on so many different biological topics. And the beauty of it however, is that by being-- I don't want to say generalists, but by being multi-specialists by working on so many different disorders and having so many different collaborations spanning all these different areas. We're actually finding surprising interconnections between them.

In particular, if you look at obesity and Alzheimer's disease, for example, there's a tremendous comorbidity between the two. So individuals who are obese are much more likely to get Alzheimer's disease. If you look at Alzheimer's and COVID-19, again, obese individuals and type 2 diabetes individuals are two-fold or more likely to die and to contact the virus.

So there's a lot of interplay between different diseases. If you now look at Alzheimer's and SARS-CoV-2, you would say, OK, great, what does one have to do with the other? Maybe obesity is the only common form here. The answer is very, very complicated. If you look at our brain, one of the things that we have been studying is the blood brain barrier. This is a set of proteins that is found in all of our vasculature, all of our vessels in the brain, the microvascular and the capillaries that basically coat them and prevent pathogens from entering the brain.

So that's the first line of defense against the brain. But every now and then pathogens do enter the brain. So what happens then if our immune system is not allowed in, then what happen? inside our brain, we have a diversity of cell types. Neurons are, of course, the stars of the show because they are the ones that transmit all of the memories, thinking, reasoning, pattern finding, et cetera. But there's a lot of cells surrounding these neurons.

So basically, there's excitatory neurons in one hand and inhibitory neurons on the other hand, both are needed to sort of create the diversity of human behaviors that we're experiencing. Oligodendrocytes and astrocytes are these large glial cells. Glial from glue, which were initially thought to just be supportive of neurons, but are found more and more to have dramatically complex functions. Oligodendrocytes basically coat the axons of neurons for electrical signals to be better insulated and better transmitted.

And when oligodendrocytes fail, then we basically lose myelin and there's a lot of cognitive loss, including in Alzheimer's disease. Now what is myelin made of its lipids? So this insulating sheath that happens surrounding the axons is made of lipids. Now in obesity, there's an enormous amount of lipid dysregulation.

So basically, if you look at insulin secretion and if you look at the phospholipid bilayer of our cells. If you look at the secretion of proteins. If you look at in X transport, what is the strongest genetic association with Alzheimer's with sporadic Alzheimer's it's ApoE apolipoprotein E with the E4 variant causing 100-fold increased risk compared to any other variant in the genome. And by far, the most important risk for Alzheimer's is ApoE4.

So what happens with the E4 version of the ApoE apolipoprotein. So what does apolipoprotein mean? So it's basically at a lipid transporter. So again, lipid front and center in Alzheimer's. So lipid transport is utilized in everyone ourselves, not just for the myelin sheath, but also for energetics. So our brain is only 3 pounds heavy. That's a tiny, tiny fraction of our overall body weight. And yet, it utilizes about 20% of our energy daily.

So this one organ is incredibly energy rich. So there have been dramatic transformations to both make our brain more energy efficient, but also to make our brain much more energy consuming compared to the rest of our body. Compared to any other primate, we have these disproportion and, of course, any other mammal fish or any other animal on Earth. We have these disproportionate allocation of energy resources in our brain.

So all glycosylation and all of these diverse processes are needed to basically generate an enormous amount of energy for our daily cognitive functions. Along with that, comes a lot of lipid regulation because lipids are a major source of energy production of ATP of all of the energetic needs of our brain. We're basically very closely related to lipids. And then there's another aspect, which is the byproducts of all of that energy production.

Every time you produce ATP, there's oxidative reactive species that are chemicals that are causing damage in your cells. So giant energetic production is also associated with cellular damage. And again, lipids are necessary for transporting all of that outside the cells and sort of clearing out the cells through the CSF, the cerebrospinal fluid through the vasculature and so on and so forth. So all of these processes are now connecting obesity with Alzheimer's. And now where do pathogens come in.

As I mentioned earlier, the immune cells of the circulating blood are not allowed. And that takes us to the next cell type of our brain, which is the microglia. So the microglia are basically the resident immune cells of our brain. So they have the same lineage as the blood circulating monocytes, and the same lineage as all of our resident immune cells of the rest of the organs known as macrophages.

So if you look at obesity, for example, M1 versus M2 macrophage in our adipose tissue will basically lead to increase or decrease risk of obesity based on, again, the role of resident immune cells in our tissues. If you look at the resident immune cells of our brain, the microglia, they played tremendous roles. One of the things that we showed back in a major paper in 2015 in collaboration with the Li-Huei Tsai is that if you look at the genetic variants associated with Alzheimer's disease, we talked about ApoE4, but there are thousands of genetic variants scattered across the genome.

If you now use our epigenomic maps, so we've constructed maps of the human epigenome for many, many years. And the reference map of the human epigenome was one of the things that we published in another issue of nature also in 2015. So as part of the epigenomics roadmap project, we basically annotated the control regions of thousands-- of hundreds of different tissues in the human body.

So looking at 117 tissues, for example, and more recently 834 tissues of the human body. We can provide a reference annotation of the control circuitry. So something extremely important to realize is that if you look at common disease, including Alzheimer's, obesity, diabetes, everything schizophrenia or psychiatric disorders. You basically have thousands of common variants that are associated with the disease. Each of which plays a very, very small role with some rare expression exceptions like FTO in obesity or ApoE in Alzheimer's.

But basically, the rest is just tiny, tiny little effects. And by far these regions are not perturbing protein coding genes, instead they're perturbing the circuitry of the cell. They're perturbative control regions that govern when genes are turned on and turned off across different tissues, different cell types, different environmental situations. So what we found in 2015 is that after having built this map of the human epigenome, we looked at brain, of course, and we looked at dozens of other tissues and cell types.

And what we found is that the genetic variants associated with Alzheimer's, we're not localizing at all in our brain signal. It was remarkable. Our brain enhancers we're just simply not lighting up at all for Alzheimer's, which was puzzling because we obviously know that it acts in the brain. But what we found is that enrichment in monocytes, and these are the resident immune cells of the brain. These are the same lineage as macrophages in other tissues. And microglia in the brain.

So that basically said, wait a minute, perhaps the microglia that only make up 5% or 10% of the cells in any one region of your brain. Perhaps, they are the culprits for why we can have an Alzheimer's predisposition. And that is also related to the amyloid hypothesis of Alzheimer's disease where in familial Alzheimer's, which is earlier onset and runs in families. The genes that are associated with Alzheimer's are pointing to all of these amyloid pathways.

But if you look at sporadic Alzheimer's you basically see a very different picture almost none of these genetic variants is in fact directly associated with amyloid. Instead these variants are localizing in these lipid transport, a lot of cardiovasculature, which we've also been studying extensively. And a lot of microglia, the immune circuitry. And now these basically points perhaps inflammation or immune processes preceding the onset of amyloid, maybe by decades.

And these microglia are also involved in the clearing out of debris, in the clearing out of all of these oxidative stress, and they are-- that's the place where ApoE, for example, gets hugely expressed astrocytes and microglia. So I mentioned oligodendrocytes are basically setting up this myelin sheath. Astrocytes are instead controlling, clearing, and sort of working with microglia serving as an interface to the environment. Touching the vasculature with their end feet and interacting with that as well.

So there's this incredibly complex interplay between all of these different processes. It sort of brings up the age old debate, I guess, can we sort of avoid this genetic determinism and what role does genetics versus environment play in human physiology and performance and cognition. Can you speak to that?

So if you look at Alzheimer's disease, you basically have this enormous role of early life in Alzheimer's. If you have an enriching environment when you are first developing, your risk of Alzheimer's in late life is decreased dramatically. If you speak a foreign language, your risk of Alzheimer's is decreased dramatically. So the environment plays a major, major role in things that happen decades later. And there's, of course, a genetic predisposition to Alzheimer's. There's a genetic predisposition to disease. There's a huge diversity of roles that genetic variants are playing.

But one of the things that we're learning from our research is that you cannot ignore the environment and you cannot ignore the genetics. Basically, there are genetic differences between humans, there's genetic differences in how our brain functions, and how our muscles function, and how our metabolism functions. For example, I'm homozygous risk for an obesity variant that predisposes me to obesity. This FTO locus that our team dissected in 2015 is the strongest genetic association with obesity. And I carry both copies of the risk earlier.

So that basically means that I have to constantly struggle to not eat, to sort of watch my diet, to watch my exercise level, in order to avoid this genetic lottery that I kind of didn't get so lucky at. So I think there's a lot to be said about the fact that every one of us has different predispositions to different types of behaviors, different types of physiology, different types of preferences, different types of disease. But at the same time, I want to bring you back to the movie *Gattaca* where the tagline was there's no gene for the human soul or something like that.

Basically, it was all about a society where decades from now, we are able to predict human physiology, predict human cognition, predict human physical performance, and use our genetic makeup to tailor people into different professions. So a dystopian future where instead of walking into an interview, they just look at your DNA and they basically say, congratulations you got the job or sorry, you didn't get the job based on your predisposition.

And again, the point of that movie was that human behavior and the human spirit can overcome a lot of the limitations that we have from our genetics. So I think that if we want an equitable society, if we want people to sort of truly flourish to the maximum that they can, we should be understanding on one end that there are genetic differences between individuals. But on the other hand, rejecting genetic determinism. So I'm not looking at my genome and saying, oh, great, I'm going to be obese, forget it.

I'm saying, well, no, this is what I got to work on. And in the same way, I know so many people who have a genetic strong effect mutation that will have doomed them for x or y or z. And they're saying, forget that. I am going to overcome the odds and I'm going to do so much more than what my genome would have quote unquote, "destined" me to. So again, as a genomicist, we are looking at the signals. We're understanding the signals, we're understanding the diversity, and the beauty of diversity in the human population.

The fact that every one of my children is so different from each other. That tells you so much about just the beauty of diversity of how our brains, our emotions functions dramatically different from birth. I mean, things that I could see at my children when they were first born. I can still see in them today many years later. So we have to embrace that every one of us is different, but also we have to embrace the fact that we are so much more than what our genome says. And we are able to overcome these shortcomings.

And sometimes, read the genome to prepare ourselves for something we might not be super happy about, but to combat it and to address it much earlier than if we didn't have that knowledge. And again, very importantly, there's a small number of strong effect mutations that every one of us carries. But by far, the vast majority of our variation is very weak effect mutations. And obesity, for example, is thought to be 70% genetic.

Intelligence is thought to be 50% genetic. Height is thought to be 70% to 80% genetic. So there's a lot of proportion of that overall phenotypic variation that is genetically encoded. But there's a big, big chunk that we have full control over based on the decisions we make, based on the choices we make, based on our nutrition, our exercise, our onset every morning. And I think that's where our educational system should be stepping in and giving opportunity to everyone to sort of really reach their full potential across the board.

That's fascinating. And hopeful at the same time. And your team, your group is doing wonderful research. We appreciate your time today, Manolis. Thank you very much for your insights. And if anybody has any questions, we'd like to learn more about the work you're doing where would you direct them.

So just email me manoli@mit.edu. You can also go on our website at compbio.mit.edu. So Computational Biology compbio.mit.edu. You will find tons of resources. We've published over 200 papers. You can find them all there. I've given dozens of talks. You can find them all there. I've recorded all of my lectures all of my MIT lectures are publicly available. You can find all of the videos of these lectures there. Every interview that I've given, including this one will be posted there. So tons of resources there, please go there.

And then you can also follow me at my Twitter page. We post all of our papers there. When I speak at different venues, I also posted there. So Manolis Kellis is just one word is my Twitter handle. So lots of information online.

If you're interested in learning more about the CSAIL Alliance program and the latest research at CSAIL, please visit our website at cap.csail.mit.edu. And listen to our podcast series on Spotify, Apple Music, or wherever you listen to your podcasts. Tune in next month for a brand new edition of the *CSAIL Alliances* podcast and stay ahead of the curve.