Betting on big data to defeat Alzheimer's disease

by Jennifer Michalowski

Denise and Sarada Fuzzell are outsiders in the Amish communities where they spend most of their days. But they've been knocking on doors here for so long, people have come to expect them. "I've heard about you," the residents often say when they answer their doors. "I've been wondering when you would come."

The visits that follow can last hours. The Fuzzells chat with their hosts about shared acquaintances, savor fresh-baked pie, and tour gardens. Sarada's even been drawn into a game of Scrabble with a local spelling bee champion. (She lost.) But ultimately, the two women come to this part of Ohio — where communication

rarely involves a telephone, let alone a computer — as part of a massive, technology-driven research effort out of Case Western Reserve University School of Medicine.

They are members of an interdisciplinary team led by genetic epidemiologist Jonathan Haines, PhD, chair of the Department of Population and Quantitative Health Sciences and the Mary W. Sheldon, MD, Professor of Genomic Sciences, who has been working with the Amish to unravel some particularly vexing genetic puzzles for nearly 20 years. They are seeking families who are willing to share their clinical information and blood samples for DNA sequencing with the team.

The Fuzzells have found that the Amish almost always agree to participate. "They're very community-oriented people," Sarada says. "If they know that they may be carrying a trait that could lead to disease, they're very interested in helping with the research to find a cure and maybe help their community down the road."

It's not hard to convince people of the potential impact of their contribution, especially when the team is seeking participants for studies about Alzheimer's disease. With at least one of every three people who live past the age of 85 expected to develop the devastating neurodegenerative disease, any information that

The Haines laboratory investigates the functional consequences of genetic variants with the goal of identifying potential drug targets.

brings researchers closer to an effective treatment or prevention strategy is of tremendous value.

Haines and his team are actively exploring the genetics behind several diseases, but since he arrived at the school in 2013 to direct Cleveland's new Institute for Computational Biology (ICB), the Alzheimer's disease component of his research program has become a major focus. Backed by a major initiative from the National Institutes of Health (NIH), researchers at the ICB and collaborators worldwide are investing heavily in obtaining and analyzing as much information as they can.

Haines thinks genome sequences from Amish families are a promising place to search for answers. The group's shared ancestry, large families, and fastidious genealogical records make it easier for his team to sift through genetic sequences to identify variants that contribute to or protect against disease. He's optimistic that his team will find clues about the biology of Alzheimer's disease in the generations' worth of data they now have from Amish families, and that their findings might suggest ways to intervene in the disease process. But the Alzheimer's field needs as many leads as it can get.

Amyloid-beta was identified as the primary component of the brain plaques that are a defining feature of Alzheimer's disease in 1984. Soon after, the protein tau was found to make up the tangled fibers that appear inside neurons in affected patients' brains. Interfering with these plaques and tangles quickly became the focus of drug development efforts. Hundreds of clinical trials have tested potential therapies, but so far these have led only to a series of high profile failures. No new drug has been approved to treat Alzheimer's disease since 2003. The five drugs that are approved for treatment of Alzheimer's disease modulate neurotransmitter signaling, and can only help manage symptoms.

"There's been a lot of work trying to use what we know to target these processes, and so far none of it has worked," Haines says.

TCGAGGGACTGAGGGACTCATTGAGCT
TATTCCAGGATAGAATCCCTACGCATACTCGAGGGACTCATTGAGCTACGATATACGTG
CTACGATATGAGGGACTCATTGAGCTACGATATGAGGGACTCATTGAGCT
CGCATACTCGAGGG MUTATIONS IN THREE DIFFERENT GENES GAGGGACT
GATATGAGATATGAGGGACTCATTGAGCTACGATATACCTACGCA

Researchers need new strategies for developing treatments that either prevent Alzheimer's disease or halt its progression, he says.

It's already becoming clear that the pathology of Alzheimer's disease involves more than just plaques and tangles, suggesting there may be better ways to intervene. There are hints that energy metabolism and inflammation contribute to the potentially decades-long process that destroys brain circuits. As researchers broaden their search, they may implicate still more players. If geneticists can identify factors that either protect against or increase the risk of Alzheimer's, scientists will have a better idea of or about which genes and pathways are likely to be involved.

To enable a comprehensive search, the genome sequences that Haines and his team obtain from participants in their own Alzheimer's studies are being pooled with thousands of others collected for Alzheimer's studies across the United States and Europe. Members of two large genetics consortia, plus three genome sequencing and analysis centers, are collaborating through the Alzheimer's Disease Sequencing Project to generate a trove of data that will be freely available to the research community, and which Haines and colleagues at the ICB have already begun mining for clues about the disease.

The project, supported by the National Institute on Aging and the National Human Genome Research Institute, is a resource-intensive effort, involving dozens of institutions and petabytes of data. Haines expects it to pay off. Thorough analysis of genomic data from tens of thousands, if not hundreds of thousands, of people may be what it takes to find solutions for this disease, he says.

The genetic factors that underlie Alzheimer's disease are undoubtedly complex. Mutations in three different genes have been linked to early-onset Alzheimer's disease, whose symptoms become apparent when affected individuals are in their thirties to mid-sixties. Just one gene, APOE, which Haines was instrumental in discovering, is known to clearly shape individuals' risk of developing the late-onset form of the disease, which is far more common. But hundreds of genetic variants likely nudge a person's risk in one direction or the other. Some of these may only be apparent when they are present in certain combinations, or when they are combined with environmental factors that remain almost entirely unknown. According to Haines, about 20 to 30 such variants have so far been linked to the late-onset disease. "We've got to try to find them all," he says.

The nature of Alzheimer's disease presents particular challenges to genetic studies. Patients suffering from dementia may not be able to give their consent to participate, and with most individuals diagnosed late in life, it's usually impossible to involve older generations in family studies. Further muddying the issue is the fact that the disease is clinically diverse and its diagnosis is not always definitive.

Funders and participants in the Alzheimer's Disease Sequencing Project are betting on the power of big data to make sense of all this. With vast improvements in both genomic and computational technologies over the past 20 years, it's now feasible to obtain, store, and analyze complete genetic information for hundreds of thousands of individuals. "The technology has gotten to a point where we can really dig into this problem in a big way," Haines says.

An initial discovery phase of the Alzheimer's Disease Sequencing Project has already generated genome sequences for thousands of individuals with and without the disease. Now the NIH is investing $24 million to support the School of Medicine and seven other academic medical centers as their researchers analyze the data over the next four years. The Consortium for Alzheimer's Sequence Analysis, a five-university group co-led by Haines, has $12.6 million in funding to identify rare genetic variants that either protect against Alzheimer's disease or increase risk.

> "There's been a lot of work trying to use what we know to target these processes, and so far none of it has worked," Haines says. Researchers need new strategies for developing treatments that either prevent Alzheimer's disease or halt its progression."

For an endeavor of this scale, the data management challenges begin long before any analysis. The Alzheimer's Disease Sequencing Project involves collaborators at institutions across the United States and Europe who are investigating Alzheimer's disease from various angles, each with a study designed to address their own specific questions. A handful of facilities carry out all the sequencing for the overall effort, but the data individual teams gather vary widely in both their content and formatting. The result is a vast collection of genome sequences and associated data that must be meticulously quality controlled and manipulated into a standardized form—meaning that after study participants' blood samples are collected, it can be years before their genome sequences reach the people who will analyze them.

GAC
TACTCGAGGGA
TCATTGAGCTACGATATA
TACGATAT

> "People are really motivated to find an answer for this disease," says Renee Laux, who coordinates the lab's efforts to enroll study participants.

A few years into the project, processes are being streamlined and the first batch of data is in. More than 500 whole genome sequences and 11,000 complete exome sequences, which cover the protein-coding parts of the genome, have already been delivered to the ICB, with thousands more expected soon. Eventually, the Alzheimer's Disease Sequencing Project expects to generate complete genome sequences — each made of up some three billion base pairs of information — for tens of thousands of individuals.

The sheer volume of data filling the ICB's servers is empowering. "I think there's a lot in there that we're going to be able to find," Haines says. But the team knows that the technology is just a tool. Finding meaning in all those sequences is going to require shrewd minds and creative approaches.

The challenge is exactly the kind of problem the ICB was created to tackle. Case Western Reserve University School of Medicine, the Cleveland Clinic Foundation, and University Hospitals Cleveland Medical Center, formed the collaborative institute in 2013 to advance scientists' ability to deepen the knowledge of biology and improve human health by taking advantage of the vast amounts of information their clinicians and researchers were already collecting. Researchers at the three institutions are working together through the ICB to share data and develop tools and infrastructure to manage and analyze it. Their work so far provides an important foundation for the analysis phase of the Alzheimer's project. "We've figured out how to deal with a lot of these data, and the next two or three years are going to be a lot of fun," Haines says.

If you ask someone at the ICB how many genomes they will need to find what they're looking for, that person is likely to tell you they see no reason to stop collecting data until they have genome sequences from everyone — every person who has developed dementia and everyone who has maintained their mental sharpness as they aged. The more data the better, everyone agrees. But they know that sequencing genomes, while cheaper than it's ever been, is still expensive — and more importantly, the need for Alzheimer's interventions is too urgent to wait. So the team is intent on getting the most out of the data they have now.

William Bush, PhD, assistant professor in the Department of Population and Quantitative Health Sciences, says real solutions will likely come by tackling the problem in many different ways. The genetic data become much richer, he says, when they are combined with additional biological information. No one is yet churning out volumes of data on gene expression or epigenetic modifications to match the DNA sequence information being compiled by the Alzheimer's consortium; but public databases and other resources are likely full of clues about the biology of the disease. Finding ways to glean the most out of available data can go a long way. "If you're clever and you put the pieces together in the right way, then you can really explain some things," Bush says.

One priority for Bush is mapping how the range of genetic variants in people with Alzheimer's disease are likely to impact the three-dimensional forms of the proteins their cells produce. He is drawing on structural information from biochemical studies and computational modeling to find protein regions where Alzheimer's-associated variants cluster. Finding sites that are commonly disrupted in affected individuals will help sort out which rare variants picked up in genome-wide analyses are relevant to the disease, he says. Pinpointing these sensitive sites also gives drug developers a head start in thinking about how to interfere with the activity of a protein's ability to spur disease progression.

Considering the genetics in context and thinking deeply about the biology of the disease is critical, team members say. The sequences and clinical records that they deal with represent real people with complicated diagnoses, says Yeunjoo Song, PhD, a computational biologist who manages incoming data for the lab. Understanding the complexity of the disease and keeping up with what others in the field are learning helps the team ask the right questions of their data and prioritize certain findings for follow-up.

TCATTGAGCTACGATATTTACGATATT
CTCATTGAGCTACGATATGAGGGACTCATTGAGCTACGATATGAGGGACTCATTGAGC
CCTACGC THE MORE DATA THE BETTER ATTGAGCT
TTCATTGAGCTACGATATGAGGGACTCATTGAGCTACGATATACCTACGCATACTCGAG
ATTGAGCTACGATATGAGGGACTCATTGAGCTACGATAT
GGGACTCATTGAGCTACGATATACGTG

Indeed, identifying variants that associate with disease is just a beginning. Five years from now, Haines hopes not just to have identified many variants that contribute to Alzheimer's disease, but also to have begun characterizing their functional effects. Across the street from the ICB's computational hub, additional members of his team work in a traditional molecular biology laboratory. The lab, whose freezers are stacked with decades' worth of DNA and blood from study participants, is equipped not just to efficiently process samples for sequence analysis, but also to investigate the functional consequences of the variants the team links to disease — paving the way toward the identification of drug targets.

It will take a wide range of expertise to get from DNA sequences to drug targets. Even though samples and sequences for the Alzheimer's Disease Sequencing Project converge on the ICB from far-flung sources, Haines thinks his team's deep knowledge of every step of the process, from study design and sample collection to functional analyses of variants, is a real asset. Team members don't work in isolation, he says, but rather understand where their data comes from and what it needs to achieve.

That breadth of experience and close working relationship within the group also keeps the team focused on the potential impact of their work. Even as each individual focuses on their piece of the pipeline, they remain aware of the disease whose devastation they are working to curb. "People are really motivated to find an answer for this disease," says Renee Laux, MS, research operations manager, who coordinates the lab's efforts to enroll study participants. She's speaking, with gratitude, about the individuals who volunteer their time and donate blood for her team's research — but the same is clearly true of her colleagues. M

Jonathan Haines, PhD