The mysterious 98%: Scientists look to shine light on the 'dark genome'

February 3 2017, by Dana Smith

Credit: David Senior

After the 2003 completion of the Human Genome Project – which sequenced all 3 billion "letters," or base pairs, in the human genome – many thought that our DNA would become an open book. But a perplexing problem quickly emerged: although scientists could transcribe the book, they could only interpret a small percentage of it.

The mysterious majority – as much as 98 percent – of our DNA do not code for proteins. Much of this "dark matter genome" is thought to be nonfunctional evolutionary leftovers that are just along for the ride. However, hidden among this noncoding DNA are many crucial regulatory elements that control the activity of thousands of genes. What is more, these elements play a major role in diseases such as cancer, heart disease, and autism, and they could hold the key to possible cures.

As part of a major ongoing effort to fully map and annotate the functional sequences of the human genome, including this silent majority, the National Institutes of Health (NIH) on Feb. 2, 2017, announced new grant funding for a nationwide project to set up five "characterization centers," including two at UC San Francisco, to study how these regulatory elements influence gene expression and, consequently, cell behavior.

The project's aim is for scientists to use the latest technology, such as genome editing, to gain insights into human biology that could one day lead to treatments for complex genetic diseases.

Importance of Genomic Grammar

After the shortfalls of the Human Genome Project became clear, the Encyclopedia of DNA Elements (ENCODE) Project was launched in September 2003 by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to find all the functional regions of the human genome, whether they form genes or not.

"The Human Genome Project mapped the letters of the human genome, but it didn't tell us anything about the grammar: where the punctuation is, where the starts and ends are," said NIH Program Director Elise Feingold, PhD. "That's what ENCODE is trying to do."
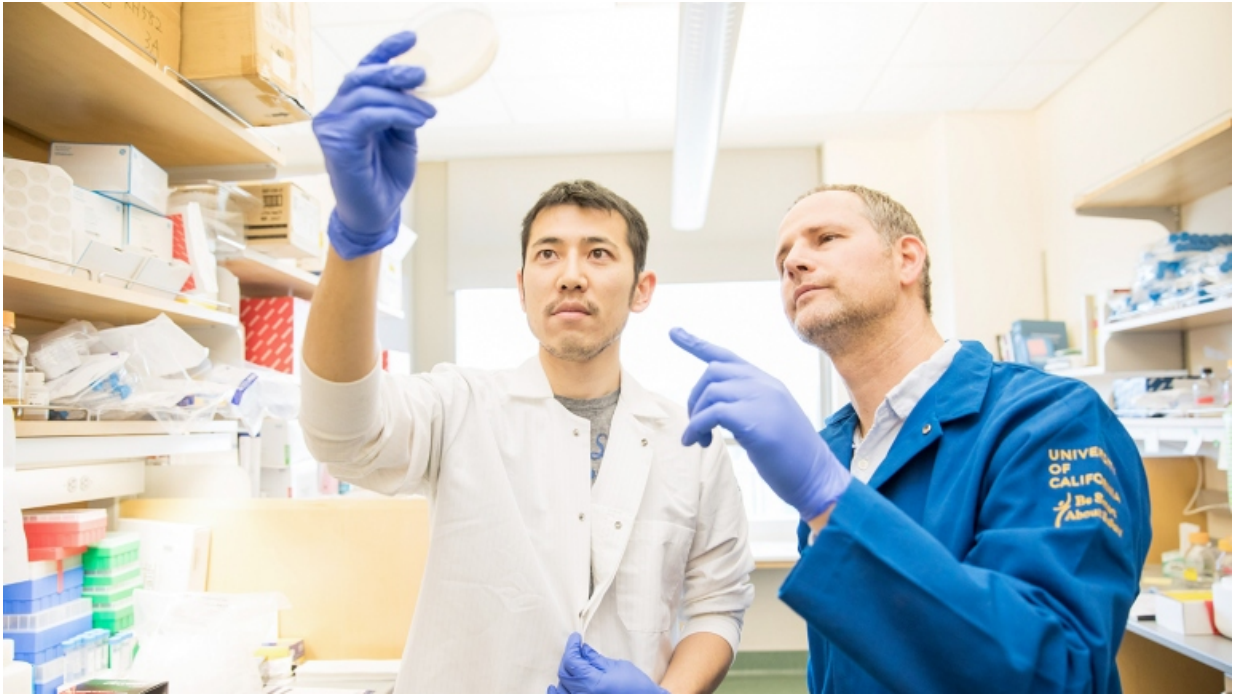
Nadav Ahituv (right), PhD, is using new technology to test for enhancers among 100,000 regulatory sequences in DNA. Credit: Susan Merrell

The initiative revealed that millions of these noncoding letter sequences perform essential regulatory actions, like turning genes on or off in different types of cells. However, while scientists have established that these regulatory sequences have important functions, they do not know what function each sequence performs, nor do they know which gene each one affects. That is because the sequences are often located far from their target genes – in some cases millions of letters away. What's more, many of the sequences have different effects in different types of cells.

The new grants from NHGRI will allow the five new centers to work to define the functions and gene targets of these regulatory sequences. At

UCSF, two of the centers will be based in the labs of Nadav Ahituv, PhD, and Yin Shen, PhD. The other three characterization centers will be housed at Stanford University, Cornell University, and the Lawrence Berkeley National Laboratory. Additional centers will continue to focus on mapping, computational analysis, data analysis and data coordination.

Cellular Barcodes Reveal Regulatory Function

New technology has made identifying the function and targets of regulatory sequences much easier. Scientists can now manipulate cells to obtain more information about their DNA, and, thanks to high-throughput screening, they can do so in large batches, testing thousands of sequences in one experiment instead of one by one.

"It used to be extremely difficult to test for function in the noncoding part of the genome," said Ahituv, a professor in the Department of Bioengineering and Therapeutic Sciences. "With a gene, it's easier to assess the effect because there is a change in the corresponding protein. But with regulatory sequences, you don't know what a change in DNA can lead to, so it's hard to predict the functional output."

Ahituv and Shen are both using innovative techniques to study enhancers, which play a fundamental role in gene expression. Every cell in the human body contains the same DNA. What determines whether a cell is a skin cell or a brain cell or a heart cell is which genes are turned on and off. Enhancers are the secret switches that turn on cell-type specific genes.

During a previous phase of ENCODE, Ahituv and collaborator Jay Shendure, PhD, at the University of Washington, developed a technique called lentivirus-based massive parallel reporter assay to identify enhancers. With the new grant, they will use this technology to test for enhancers among 100,000 regulatory sequences previously identified by
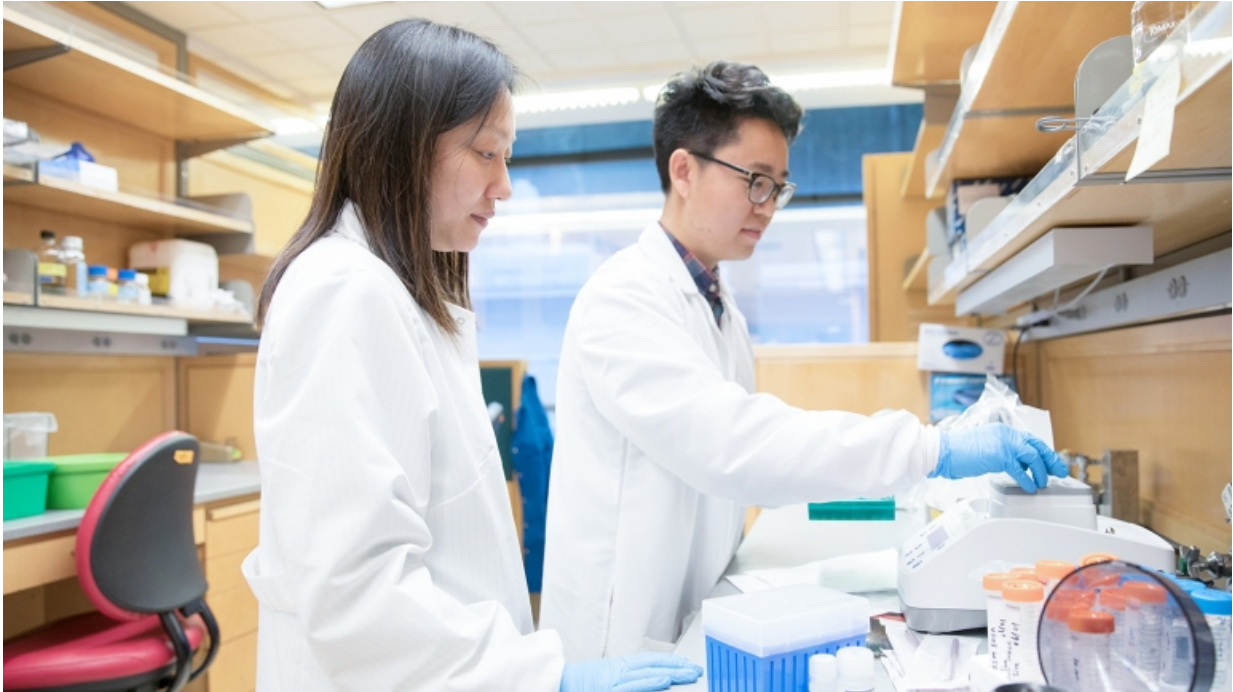
Yin Shen (left), PhD, will use CRISPR to to identify not only which noncoding sequences of DNA have regulatory functions, but also which genes they affect. Credit: Susan Merrell

Their approach pairs each regulatory sequence with a unique DNA barcode of 15 randomly generated letters. A reporter gene is stuck in between the sequence and the barcode, and the whole package is inserted into a cell. If the regulatory sequence is an enhancer, the reporter gene will turn on and activate the barcode. The DNA barcode will then code for RNA in the cell.

Once the researchers see that the reporter gene is turned on, they can easily sequence the RNA in the cell to see which barcode is activated.

They then match the barcode back to its corresponding regulatory sequence, which the scientists now know is an enhancer.

"With previous enhancer assays, you had to test each sequence one by one," Ahituv explained. "With our approach, we can clone thousands of sequences along with thousands of barcodes and test them all at once."

Deleting Sequences to Understand Their Role

Shen, an assistant professor in the Department of Neurology and the Institute for Human Genetics, is taking a different approach to characterize the function of regulatory sequences. In collaboration with her former mentor at the Ludwig Institute for Cancer Research and UC San Diego, Bing Ren, PhD, she developed a high-throughput CRISPR-Cas9 screening method to test the function of noncoding sequences. Now, Shen and Ren are using this approach to identify not only which sequences have regulatory functions, but also which genes they affect.

Shen will use CRISPR to edit tens of thousands of regulatory sequences in a large pool of cells and track the effects of the edits on a set of 60 pairs of genes that commonly co-express.

For this work, each cell will be programmed to reflect two fluorescent colors – one for each gene – when a pair of genes is turned on. If the light in a cell goes out, the scientists will know that its target gene has been affected by one of the CRISPR-based sequence edits. The final step is to sequence each cell's DNA to determine which regulatory sequence edit caused the change in gene expression.

The causes of common diseases such as diabetes, cancer and autism are not just a gene that is changed, but the regulatory sequence in human DNA that regulates that gene, studies have shown. Credit: University of California, San Francisco

By monitoring the colors of co-expressed genes, Shen will reveal the complex relationship between numerous functional sequences and multiple genes, which was beyond the scope of traditional sequencing techniques.

"Until the recent development of CRISPR, it was not possible to genetically manipulate non-coding sequences in a large scale," said Shen. "Now, CRISPR can be scaled up so that we can screen thousands of regulatory sequences in one experiment. This approach will tell us not only which sequences are functional in a cell, but also which gene they regulate."

Can Dark Matter DNA Treat Disease?

By cataloging the functions of thousands of regulatory sequences, Shen and Ahituv hope to develop rules about how to predict and interpret other sequences' functions. This would not only help illuminate the rest of the dark matter genome, it could also reveal new treatment targets for complex genetic diseases.

"A lot of human diseases have been found to be associated with regulatory sequences," Ahituv said. "For example, in genome-wide association studies for common diseases, such as diabetes, cancer and autism, 90 percent of the disease-associated DNA variants are in the noncoding DNA. So it's not a gene that's changed, but what regulates it."

As the price for sequencing a person's genome has dropped significantly, there is talk about using precision medicine to cure many serious diseases. However, the hurdle of how to interpret mutations in noncoding DNA remains.

"If we can characterize the function and identify the gene targets of these regulatory sequences, we can start to reveal how their mutations contribute to diseases," Shen said. "Eventually, we may even be able to treat complex diseases by correcting regulatory mutations."

Provided by University of California, San Francisco