# Bits of Mystery DNA, Far From 'Junk,' Play Crucial Role

By **GINA KOLATA**

Among the many mysteries of human biology is why complex diseases like diabetes, high blood pressure and psychiatric disorders are so difficult to predict and, often, to treat. An equally perplexing puzzle is why one individual gets a disease like cancer or depression, while an identical twin remains perfectly healthy.



Béatrice de Géa for The New York Times

"It is like opening a wiring closet and seeing a hairball of wires," Mark Gerstein of Yale University said of the DNA intricacies.

Now scientists have discovered a vital clue to unraveling these riddles. The human genome is packed with at least four million gene switches that reside in bits of DNA that once were dismissed as "junk" but that turn out to play critical roles in controlling how cells, organs and other tissues behave. The discovery, considered a major medical and scientific breakthrough, has enormous implications for human health because many complex diseases appear to be caused by tiny changes in hundreds of gene switches.

The findings, which are the fruit of an immense federal project involving 440 scientists from 32 laboratories around the world, will have immediate applications for understanding how alterations in the non-gene parts of DNA contribute to human diseases, which may in turn lead to new drugs. They can also help explain how the environment can affect disease risk. In the case of identical twins, small changes in environmental exposure can slightly alter gene switches, with the result that one twin gets a disease and the other does not.

As scientists delved into the "junk" — parts of the DNA that are not actual genes containing instructions for proteins — they discovered a complex system that controls genes. At least 80 percent of this DNA is active and needed. The result of the work is an annotated road map of much of this DNA, noting what it is doing and how. It includes the system of switches that, acting like dimmer switches for lights, control which genes are

used in a cell and when they are used, and determine, for instance, whether a cell becomes a liver cell or a neuron.

"It's Google Maps," said Eric Lander, president of the Broad Institute, a joint research endeavor of Harvard and the Massachusetts Institute of Technology. In contrast, the project's predecessor, the Human Genome Project, which determined the entire sequence of human DNA, "was like getting a picture of Earth from space," he said. "It doesn't tell you where the roads are, it doesn't tell you what traffic is like at what time of the day, it doesn't tell you where the good restaurants are, or the hospitals or the cities or the rivers."

The new result "is a stunning resource," said Dr. Lander, who was not involved in the research that produced it but was a leader in the Human Genome Project. "My head explodes at the amount of data."

The discoveries were published on Wednesday in six papers in the journal Nature and in 24 papers in Genome Research and Genome Biology. In addition, The Journal of Biological Chemistry is publishing six review articles, and Science is publishing yet another article.

Human DNA is "a lot more active than we expected, and there are a lot more things happening than we expected," said Ewan Birney of the European Molecular Biology Laboratory-European Bioinformatics Institute, a lead researcher on the project.

In one of the Nature papers, researchers link the gene switches to a range of human diseases — multiple sclerosis, lupus, rheumatoid arthritis, Crohn's disease, celiac disease — and even to traits like height. In large studies over the past decade, scientists found that minor changes in human DNA sequences increase the risk that a person will get those diseases. But those changes were in the junk, now often referred to as the dark matter — they were not changes in genes — and their significance was not clear. The new analysis reveals that a great many of those changes alter gene switches and are highly significant.

"Most of the changes that affect disease don't lie in the genes themselves; they lie in the switches," said Michael Snyder, a Stanford University researcher for the project, called Encode, for Encyclopedia of DNA Elements.

And that, said Dr. Bradley Bernstein, an Encode researcher at Massachusetts General Hospital, "is a really big deal." He added, "I don't think anyone predicted that would be the case."

The discoveries also can reveal which genetic changes are important in cancer, and why. As they began determining the DNA sequences of cancer cells, researchers realized that most of the thousands of DNA changes in cancer cells were not in genes; they were in the dark matter. The challenge is to figure out which of those changes are driving the cancer's growth.

"These papers are very significant," said Dr. Mark A. Rubin, a prostate cancer genomics researcher at Weill Cornell Medical College. Dr. Rubin, who was not part of the Encode project, added, "They will definitely have an impact on our medical research on cancer."

In prostate cancer, for example, his group found mutations in important genes that are not readily attacked by drugs. But Encode, by showing which regions of the dark matter control those genes, gives another way to attack them: target those controlling switches.

Dr. Rubin, who also used the Google Maps analogy, explained: "Now you can follow the roads and see the traffic circulation. That's exactly the same way we will use these data in cancer research." Encode provides a road map with traffic patterns for alternate ways to go after cancer genes, he said.

Dr. Bernstein said, "This is a resource, like the human genome, that will drive science forward."

The system, though, is stunningly complex, with many redundancies. Just the idea of so many switches was almost incomprehensible, Dr. Bernstein said.

There also is a sort of DNA wiring system that is almost inconceivably intricate.

"It is like opening a wiring closet and seeing a hairball of wires," said Mark Gerstein, an Encode researcher from Yale. "We tried to unravel this hairball and make it interpretable."

There is another sort of hairball as well: the complex three-dimensional structure of DNA. Human DNA is such a long strand — about 10 feet of DNA stuffed into a microscopic nucleus of a cell — that it fits only because it is tightly wound and coiled around itself. When they looked at the three-dimensional structure — the hairball — Encode researchers discovered that small segments of dark-matter DNA are often quite close to genes they control. In the past, when they analyzed only the uncoiled length of DNA, those controlling regions appeared to be far from the genes they affect.

The project began in 2003, as researchers began to appreciate how little they knew about human DNA. In recent years, some began to find switches in the 99 percent of human DNA that is not genes, but they could not fully characterize or explain what a vast majority of it was doing.

The thought before the start of the project, said Thomas Gingeras, an Encode researcher from Cold Spring Harbor Laboratory, was that only 5 to 10 percent of the DNA in a human being was actually being used.

The big surprise was not only that almost all of the DNA is used but also that a large proportion of it is gene switches. Before Encode, said Dr. John Stamatoyannopoulos, a University of Washington scientist who was part of the project, "if you had said half of

the genome and probably more has instructions for turning genes on and off, I don't think people would have believed you."

By the time the National Human Genome Research Institute, part of the National Institutes of Health, embarked on Encode, major advances in DNA sequencing and computational biology had made it conceivable to try to understand the dark matter of human DNA. Even so, the analysis was daunting — the researchers generated 15 trillion bytes of raw data. Analyzing the data required the equivalent of more than 300 years of computer time.

Just organizing the researchers and coordinating the work was a huge undertaking. Dr. Gerstein, one of the project's leaders, has produced a diagram of the authors with their connections to one another. It looks nearly as complicated as the wiring diagram for the human DNA switches. Now that part of the work is done, and the hundreds of authors have written their papers.

"There is literally a flotilla of papers," Dr. Gerstein said. But, he added, more work has yet to be done — there are still parts of the genome that have not been figured out.

That, though, is for the next stage of Encode.