**ENCODE Exposure?**

**Newly Revealed Genome Complexity May Mean Personalized Medicine Is Farther Away**

By Charlie Schmidt

A global effort to find all the functional elements in DNA has revealed surprising genomic complexity that will probably complicate progress toward personalized medicine.

The findings suggest that enormous regions of the genome once dubbed “junk DNA” are functional, even though they don’t code for proteins. The discovery—described in *Nature* in June and detailed in 28 companion reports in the June issue of *Genome Research*—was made by the ENCODE (Encyclopedia of DNA Elements) consortium, a collection of 35 research groups from around the world headed by the National Human Genome Research Institute (NHGRI) at the National Institutes of Health.

ENCODE scientists exhaustively studied 30 million base pairs of human DNA, about 1% of the genome, and found evidence of transcription in the regions that lie between individual genes. (Transcription is the process by which DNA makes RNA, the molecule from which proteins are generated.) Scientists used to think that only the protein-coding components of genes—making up just 2% of the genome at most—were involved in transcription. The rest of the genome, they assumed, was largely inert.

ENCODE’s findings have shown this assumption to be false. However, precisely what the noncoding regions do remains a mystery, concedes Eric Green, M.D., Ph.D., NHGRI’s scientific director. “We understand the language of genes and the genetic code, but I can’t stress enough how little we understand about the language of noncoding DNA,” he says. “Protein production is governed by straightforward rules, but noncoding DNA appears to have multiple functions, each coordinated by different rules. We’ve only scratched the surface in terms of our understanding of how the genome works.”

Far from being a collection of tidy genes with independent effects, the genome is a vast network of genes and other regulatory elements that interact in unknown ways. As the reality of this complexity sets in, scientists are revising expectations for personalized medicine and setting projected timelines back indefinitely.

“In the short term, it makes everything more difficult,” asserts John Greally, M.D., Ph.D., an associate professor of medicine and molecular genetics at Albert Einstein College of Medicine. “Instead of looking only in genes for disease variations, we have to consider the entire genome. But it’s only with more complete information that we can hope to understand the whole system.”

After decoding the human genome, scientists predicted that progress toward personalized medicine would advance rapidly. But despite 25,000 research reports since 2000 investigating disease associations with variations in nearly 2,500 genes, researchers have yet to identify any new variations with substantial clinical effects, according to Muin Khoury, M.D., Ph.D., who directs the National Office of Public Health Genomics at the Centers for Disease Control and Prevention. At most, 10% of published associations have been validated by other researchers, Khoury says. And few of these, if any, have resulted in clinical or public health applications.

Of particular significance is the lack of any new “big genes with big effects,” such as the BRCA gene variations that raise breast cancer risk by up to 85%. Women who inherit the BRCA variations can benefit from personalized care, including prophylactic surgery. Scientists hoped that the genome would yield more genes like BRCA, which would point to clear clinical opportunities. But instead, genomic research has consistently linked illnesses with networks of smaller genes, each with minor risk contributions that vary greatly from patient to patient. Validating the effects of these genes is difficult, in part because enormous numbers of patients are needed to boost statistical power enough to make these small changes visible. Thus, with few exceptions, gene networks have not generated drug targets or clinically useful biomarkers.

**Complexity Poses New Challenges**

Now, researchers must also contend with a vast universe of noncoding DNA that they know virtually nothing about. Although he’s reluctant to predict timelines, Green suggests that the functionality of variants in noncoding regions could take 4–5 years longer to assess than those found in genes.

“If you give me a variant in a coding region, I can get clues about the likelihood of its being functionally important quickly,” he
says. “That's not true for variants in non-coding regions. We don’t know anything about how they’re packaged, transcribed, or regulated.”

Scientists say they weren’t entirely surprised by the ENCODE findings. For one thing, the notion that 98% of the genome has no useful purpose doesn’t make intuitive sense, they say. What’s more, evidence that intergenic regions regulate gene behavior had already begun to emerge. For instance, the gene that codes for globin protein, which makes hemoglobin for red blood cells, is regulated by locus control regions located beyond the gene itself, a fact that has been understood since the mid-1990s.

But it wasn’t until ENCODE produced its results that scientists realized how pervasive intergenic transcription really is. Over 4 years, scientists working with the ENCODE consortium used high-throughput tools to scour DNA for functional elements, including protein-coding genes, genes that don’t code for proteins, regulatory elements that control gene transcription, and elements that maintain the structure and replication of chromosomes. That research ultimately showed that up to 93% of the genome is transcribed at some point.

Nongene sequences were assumed to be functional, Greally explains, if they had chromatin marks, structural changes that reflect RNA transcription. Researchers also assumed that a region was functional if the sequences were highly conserved or replicated throughout the genome. According to Thomas Gingeras, Ph.D., vice president for biological research at Affymetrix Inc., the results indicate that the entire genome, including intergenic regions and genes combined, generates two key products: proteins, which perform the cell’s catalytic activities, and RNA, which has a wide range of possible functions. For instance, microRNAs and small interfering RNAs are increasingly known to crucially regulate gene expression, he says.

What ENCODE can offer, Green explains, is an opportunity to see where nongene sequences are located in the DNA molecule. That ability will allow scientists to infer which of the variants they identify during disease-association studies are biologically important. The growing ENCODE database resides on a Web portal hosted by the University of California, Santa Cruz (http://genome.ucsc.edu/ENCODE/).

“Say you find a sequence variant in non-coding DNA that shows up more often in people with a particular cancer than it does in healthy people,” Greally explains. “By screening the variant in ENCODE, you might find it exists in the middle of a transcriptional element that seems to be important to a gene. And that might give you a clue about how it fits into a particular biochemical pathway.”

Mary Relling, Pharm.D., chair of the pharmaceutical department at St. Jude’s Children’s Hospital in Memphis, Tenn., says she applauds ENCODE’s mission to identify functional elements in noncoding regions. “Before I spend thousands of dollars and a postdoc’s lifetime investigating some obscure intergenic polymorphism, I want evidence that it has a functional consequence,” she says. “And if it exists in a region that’s already been studied by ENCODE, then I don’t have to look for that evidence myself. It means we don’t have to reinvent the wheel.”

But Relling admits to being daunted by ENCODE’s conclusions. “The thing that’s worrisome is that there are many mechanisms by which noncoding DNA could work. It’s going to take an unbelievable effort to determine what it does.”

**Reality Versus Expectations**

Meanwhile, despite growing evidence that most inherited diseases arise from complex genetic networks, the media continues to report genomic findings in gene-centric ways. Hardly a day goes by without news of some newly discovered gene for a given disease that almost certainly hasn’t been validated in several studies. These reports fuel public expectations that personalized medicine is more advanced than it is. Indeed, the ENCODE findings—while crucial for fundamental research—illustrate just how far genomic science has to go before it radically transforms medical care.

Relling concedes that researchers didn’t sufficiently consider the numbers of patients needed to test genomic associations with clinical outcomes. That’s particularly true for outcomes driven by multiple genetic variations working in combination, she says. Assessing the unique contributions of those combinations requires studies in tens of thousands to hundreds of thousands of people.

Given those constraints, it makes more sense to view the current state of genomic research as a time of discovery than a period of rapid clinical change, argues Christine Ambrosone, Ph.D., a professor of oncology at Roswell Park Cancer Institute in Buffalo, N.Y. “We’re just doing the basic science,” she says. “It may not have an immediate effect on personalized medicine, but it’s opening doors to more research and investigation that I hope will get us there. This is an incredibly exciting time to be in research; new discoveries are being made constantly.”

The pace of progress, researchers say, will depend largely on the costs of analytical platforms for detecting DNA variations and technology advancements. Indeed, the high cost and limited availability of genomewide sequencing might explain why additional genes (or nongenes) with big effects haven’t yet been identified, Relling suggests. Green says technology improvements that allow researchers to scale up their investigations to a genomewide level are now a key priority for ENCODE. In future efforts, project scientists will also focus on exploring functionality in the 1% of the genome analyzed so far, even as efforts to study the remaining 99% move forward, he says.

According to Gingeras, an NHGRI program called Full Genome ENCODE Analysis, which will run for 4 years beginning this month, will coordinate upcoming
The program had not yet been formally announced when this article went to press. The research will focus in part on functional operations in noncoding regions in the cell, Gingeras says, and ideally it will expand its investigations to other cell types. For practical purposes, the ENCODE project focused on just a few human cell lines—namely, HeLa, an established line of human cervical cancer cells, and HL60 cells, a line of human leukemia cells—that can be easily cultured and distributed for research. But Greally cautions that these cell lines have critical shortcomings, such as broken chromosomes and unusual additions to and losses of DNA. “So, whether the regulatory processes in these cell lines are representative of those in primary cells from the human body is, to say the least, questionable,” he wrote in an editorial that appeared with the group ENCODE report in *Nature*. Of course, the notion that the nongene regulatory processes they have identified are different for different cell types—which is almost certainly true—merely adds to the research burden. To fully accomplish its mission, ENCODE will have to study the entire genome of every cell type from blood to bone to brain—a daunting task.

What’s more, scientists increasingly recognize that inherited variations in both gene and nongene regions are further influenced by epigenetic changes to DNA that occur after conception. Thus, while the glimpse into noncoding regions achieved thus far is tantalizing, it merely confirms the magnitude of the challenge that looms ahead, Greally emphasizes.

Ultimately, the genomic complexity that ENCODE revealed gives researchers a new way to think about cancer and other diseases. More than interplay of genes and environment, cancer probably arises from an interplay of environmental factors and a wide range of disturbances in gene regulation—occurring not just at the single-gene level but throughout the genome.

Contemplating that prospect, David Hafler, M.D., a professor of neurology at Harvard Medical School, strikes a philosophical tone: “You have to consider that all this will probably take generations to sort out and work through,” he says. “My sense is that things are starting to happen; it may be slower than we’d like, and progress so far might not match our initial expectations. We have to accept that the variants we’ve been finding—and I think we could have predicted this—tend to have minor effects. Whether they have any clinical uses, only time will tell.”

© Oxford University Press 2007. DOI: 10.1093/jnci/djm220