

FEBRUARY 12, 2001

Human Genome Analysis Hints at New Proteins Involved in Gene Expression

An early search of the draft human genome sequence has revealed promising evidence that the completed genome will yield new proteins involved in gene expression. The researchers who performed the analysis found previously unknown genes that appear to encode proteins involved in gene expression.

The researchers also found evidence that the human gene expression machinery is more complex than that of lower animals. This discovery suggests that the more sophisticated machinery for translating genetic information "may be particularly important for shaping human development and physiology," wrote the researchers.

"This search illustrates how the human genome sequence will provide many new factors that may be involved in gene expression, although their roles remain unknown."

— **Michael R. Green**

Howard Hughes Medical Institute investigator Michael R. Green at the University of Massachusetts Medical School and co-authors Rossella Tupler at the Università degli Studi di Pavia and Giovanni Perini at the University of Bologna published their analysis in the February 15, 2001, issue of the journal *Nature*. The analysis is part of a collection of papers published by *Nature* that discusses the implications of efforts to sequence the human genome.

"The availability of the human genome and other genome sequences will revolutionize all fields of biomedical research," wrote the scientists. "But, as the genome itself is the object of gene expression, the impact may be particularly profound for those of us studying this process."

In an interview discussing the *Nature* article, Green added, "in studying gene expression, as opposed to most other biological processes it is the genome itself that is being explored. The genome contains both the sequences of the players involved—the proteins—and the sequences of all the signals that

govern the expression of the genes."

In their analyses, the scientists searched for genes governing three steps in gene expression—the transcription of the gene into messenger RNA (mRNA); the early splicing of mRNA to produce the final molecule that will specify a protein; and the addition of the poly(A) "tail" to one end of the mRNA that is necessary for its processing by the protein machinery.

In searching the human genome for gene sequences that specify general transcription factors (GTFs), and transcriptional activators, the researchers found sequences for numerous genes that were similar to those in yeast and the fruitfly *Drosophila*. However, they also found many more human gene sequences than in the *Drosophila* genome that appear to be related to a particular GTF, called TFIID, "indicating that the potential diversity of human TFIID is much greater than that of *Drosophila*."

In searching for transcriptional activators, the researchers found more than 2,000 genes that could code for these proteins—far more than they found in genome databases for *Drosophila* and the roundworm *C. elegans*. "This search illustrates how the human genome sequence will provide many new factors that may be involved in gene expression, although their roles remain unknown," wrote the scientists.

Similarly in a search for genes that encode proteins involved in assembling the mRNA splicing machinery, the scientists found "significantly greater complexity than is found in *Drosophila*." And, in searching for genes that encode proteins involved in adding the poly(A) tail to mRNA, the researchers unexpectedly found new genes, which, again, suggests that humans have a more complex gene expression machinery.

Green said, "in each case, the surprises are of the same type. When we looked for genes for certain types of proteins that were homologs to *Drosophila*, there was no reason to believe there would be more than a single protein. But in fact, we found homologs that there was no reason to suspect would be there."

The researchers pointed out, however, that "although these searches highlight the power of the new genomic information, they also reveal important limitations. In particular, the existence of a related gene does not mean that there is a corresponding protein: the sequence could be a non-expressed pseudogene."

Green explained that pseudogenes are "fossils" of real genes. "But the genome hasn't expressed them," he said. "It has kept them there, but they don't do anything."

"I suspect that many of these new genes are not pseudogenes. In some cases we can find them in databases of expressed genes. And, they are conserved to a higher degree than I would have predicted if they were pseudogenes. If they were just not expressed, there would be no evolutionary constraint to maintain sequence conservation."

The findings raise a number of intriguing questions that will take time to sort out, says Green. "These findings are very interesting and exciting, but, of course, now we really have to figure out what they mean in the context of gene expression. Are these additional proteins made during certain times of development? In certain cells? What are they doing? What are their target genes? But we wouldn't be in a position to ask these questions if we didn't know these candidates existed."

Many of the new genes appear to express subunits of larger protein complexes. Further study will be needed to understand how they are involved in those complexes. "As with any discovery, it opens up a whole series of new questions and approaches that we can explore," Green said.