DNA Study Forces Rethink of What It Means to Be a Gene

Genes, move over. Ever since the early 1900s, biologists have thought about heredity primarily in terms of genes. Today, they often view genes as compact, information-laden gems hidden among billions of bases of junk DNA. But genes, it turns out, are neither compact nor uniquely important. According to a painstaking new analysis of 1% of the human genome, genes can be sprawling, with far-flung protein-coding and regulatory regions that overlap with other genes.

As part of the Encyclopedia of DNA Elements (ENCODEx) project, 35 research teams have analyzed 44 regions of the human genome covering 30 million bases and figured out how each base contributes to overall genome function. The results, compiled in a paper in the 14 June issue of Nature and 28 papers in the June issue of Genome Research, provide a litany of new insights and drive home how complex our genetic code really is. For example, protein-coding DNA makes up barely 2% of the overall genome, yet 80% of the bases studied showed signs of being expressed, says Ewan Birney of the European Molecular Biology Laboratory’s European Bioinformatics Institute in Hinxton, U.K., who led the ENCODE analysis.

Given the traditional gene-centric perspective, that finding “is going to be very disturbing to some people,” says John Greally, a molecular biologist at Albert Einstein College of Medicine in New York City. On the other hand, not just genes but also regulatory DNA and other important features such as gene start sites. NHGRI now plans to spend about $23 million annually over the next 4 years to perform a similar analysis of the whole genome, expecting that the lessons learned from the pilot ENCODE and new sequencing technologies will greatly reduce the costs of this extended project (Science, 25 May, p. 1120). “The goal is to measure all the different kinds of features across the human genome and ask which features go together to understand the whole package,” says George Weinstock, a geneticist at Baylor College of Medicine in Houston, Texas.

A key component of the pilot ENCODE is an analysis of the “transcriptome,” the repertoire of RNA molecules that cells create by transcribing DNA. For protein-coding genes, most of their RNA transcripts—the messenger RNA (mRNA)—get translated into chains of amino acids by ribosomes. For other types of “genes,” RNA is the end product: Ribosomal RNAs become the backbone of the ribosome, for example.

Researchers used to think very little RNA was produced beyond mRNA and a smattering of RNA end products. But about half the transcripts that molecular biologist Thomas Gingeras of Affymetrix Inc. in Santa Clara, California, discovered in his RNA survey 2 years ago didn’t fit into these categories (Science, 20 May 2005, p. 1149), a finding ENCODE has now substantiated. The ENCODE researchers knew going in that the DNA they were studying produced about eight non-protein-coding RNAs, and they have now discovered thousands more. “A lot more of the DNA [is] turning up in RNA than most people would have predicted,” says Collins.

ENCODEx has produced few clues as to what these RNAs do—leaving some to wonder whether experimental artifacts inflated the percentage of DNA transcribed. Greally is satisfied that ENCODE used enough different techniques to show that the RNA transcripts are real, but he’s not sure they’re biologically important. “It’s possible some of these transcripts are just the polymerase [enzyme] chugging along like an Energizer bunny” and transcribing extra DNA, he suggests.

But in the 8 June issue of Science (p. 1484) Gingeras and his colleagues reported that many of the mysterious RNA transcripts found as part of ENCODE harbor short sequences, conserved across mice and humans, that are likely important in gene regulation. That these transcripts are “so diverse and prevalent across the genome just opens up the complexity of this whole system,” says Gingeras.
The mRNA produced from protein-coding genes also held surprises. When Alexandre Reymond, a medical geneticist at the University of Lausanne, Switzerland, and his colleagues took a close look at the 400 protein-coding genes contained in ENCODE’s target DNA, they found additional exons—the regions that code for amino acids—for more than 80%. Many of these newfound exons were located hundreds of bases away from the gene’s previously known exons, sometimes hidden in another gene. Moreover, some mRNAs were derived from exons belonging to two genes, a finding, says Reymond, that “underscores that we have still not truly answered the question, ‘What is a gene?’” In addition, further extending and blurring gene boundaries, ENCODE uncovered a slew of novel “start sites” for genes—the DNA sequences where transcription begins—many located hundreds of thousands of bases away from the known start sites.

Before ENCODE started, researchers knew of about 532 promoters, regulatory DNA that helps jump-start gene activity, in the human DNA chosen for analysis. Now they have 775 in hand, with more awaiting verification. Unexpectedly, about one-quarter of the promoters discovered were at the ends of the genes instead of at the beginning.

The distributions of exons, promoters, gene start sites, and other DNA features and the existence of widespread transcription suggest that a multidimensional network regulates gene expression. Gingeras contend that because of this complexity, researchers should look at RNA transcripts and not genes as the fundamental functional units of genomes. But Collins is more circumspect. The gene “is a concept that’s not going out of fashion,” he predicts. “It’s just that we have to be more thoughtful about it.”

—ELIZABETH PENNISI

### SYNTHETIC BIOLOGY

**Attempt to Patent Artificial Organism Draws a Protest**

An activist group’s concern about maverick genome sequencer J. Craig Venter’s intention to patent an entirely synthetic free-living organism has thrown a spotlight on the emerging intellectual-property landscape in this hot new field. The protesters claim that Venter wants his company to become the Microsoft of synthetic biology, dominating the industry.

Venter hopes to use the artificial life form, which he says does not yet exist, as a carrier for genes that would enable the bug to crank out hydrogen or ethanol to produce cheap energy. Duke University law professor Arti Rai says the patent, if awarded, “could be problematic” only if Venter’s product became the standard in the field. But Venter says this application is just the start: He plans to patent methods that would cover more than the single microbe described in the application. “We’d certainly like the freedom to operate on all synthetic organisms” that could serve as a chassis for swapping out genes, says Venter, whose research team is at the nonprofit J. Craig Venter Institute in Rockville, Maryland, but who recently started a company to commercialize the work.

Filed last October and published by the U.S. Patent and Trademark Office on 31 May, the application describes “a minimal set of protein-coding genes which provides the information required for replication of a free-living organism in a rich bacterial culture medium.” The application cites work by Hamilton Smith and others on Venter’s team on a simple bacterium called *Mycoplasma genitalium* that they are using to determine the minimum number of genes for life. They want to synthesize this “minimal genome” from scratch, get it working inside a cell, then add genes to produce cheap fuels (*Science*, 14 February 2003, p. 1006).

In a press release, the ETC Group, a technology watchdog in Ottawa, Canada, called Venter’s “monopoly claims … the start of a high-stakes commercial race to synthesize and privatize synthetic life forms.” ETC calls for the U.S. and international patent offices to reject the patent so that societal implications can be considered. ETC also cited a recent *Newsweek* interview in which the scientist says he wants to create “the first billion- or trillion-dollar organism.”

Venter says this is just one of several patent applications that would give his company, Synthetic Genomics Inc., exclusive rights to methods for making synthetic organisms. The artificial *Mycoplasma* “may or may not be” the one used to generate hydrogen or ethanol, he says; his team is working on several species. “We haven’t given any thought to the licensing conditions, but in any case, they would not impede work in academic labs,” says Venter, adding, “This is a problem that we hope will have hundreds of solutions.”

Rai says the notion that Venter’s *Mycoplasma* strain will dominate the way Microsoft’s Windows did is tenuous because “about 10 things would have to happen,” among them that Venter would create the organism, get the patent, and others would adopt his technology as the standard. Even if that happened, Venter “could do well [financially] and do good,” she says, by licensing the technology at low cost as a research tool, as happened with the original patents on recombinant DNA technology.

Other synthetic biologists don’t seem fazed. “He’s shooting an arrow in the general direction that things are going,” says Frederick Blattner of the University of Wisconsin, Madison, who has patented a stripped-down *Escherichia coli* and founded a company called Scarab Genomics that is commercializing the technology while disbursing it to academic researchers for a small cost. The more pertinent question, says Harvard’s George Church, is whether the inventors’ claims to have devised something useful will hold up, as there’s no obvious reason why a completely synthetic *Mycoplasma* is needed rather than, say, modified *E. coli* to make hydrogen.

Massachusetts Institute of Technology synthetic biologist Tom Knight, who has pointed out that anyone could get around the patent simply by adding more than the 450 genes stipulated, says his complaint is that the application doesn’t explain how to build the artificial cell. “I think it’s rather tasteless,” Knight says.

—JOCELYN KAISER