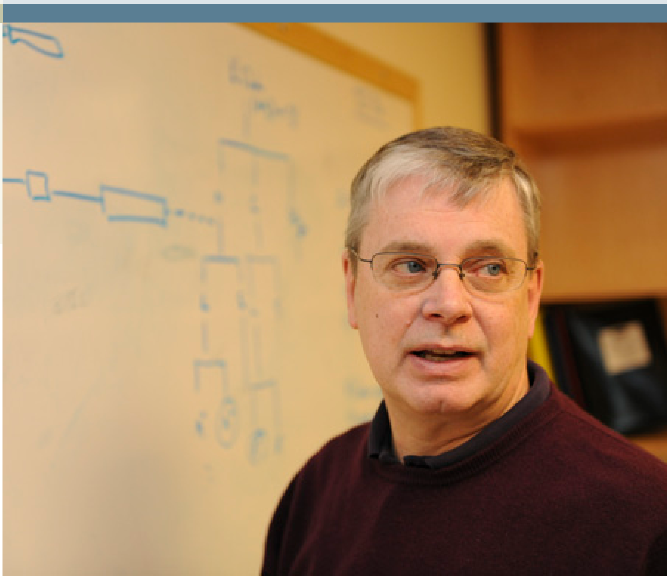
The gene, redefined

"Almost every part of the genome has a function, although in many cases we don't know the right context in which to appreciate what that function is."



A half-century ago, scientists in the young field of molecular biology figured they had a pretty good notion of how the genetic code operated. Back then, decades before the advent of genome sequencing, the human genome's 23 chromosomes were thought to harbor as many as one million distinct genes — each presumed to encode a single molecule of RNA, the template, it was then believed, for the synthesis of one protein. Proteins were understood to be the basis of most functions in the cell, including the regulation of genes themselves.

Fast-forward to 2001, the year in which biologists and computer scientists pieced together a draft sequence of the 3 billion pairs of chemical "bases," or nucleotides, that comprise the human genome. Among the first surprises was a preliminary gene count. There were not a million, not 100,000, not even 50,000; in the emerging consensus it appeared that the genome of *H. sapiens* contained fewer than 25,000 genes.

How could 25,000 genes give rise to a million distinct human proteins? Alas, the "one gene-one protein" orthodoxy had long since been overturned and replaced by a concept called alternative splicing, which explained how a single gene could generate many different RNA "messengers" and potentially, therefore, multiple proteins.

Alternative splicing is one of many phenomena that have complicated our notion of how the genome is organized and how its many elements function. Not long after the assembly of the draft human genome, a public research consortium called ENCODE (The Encyclopedia of DNA Elements) was launched by the National Human Genome Research Institute. Its aim: to compile a comprehensive list of functional elements across the human genome.

Tom Gingeras and the ENCODE "heresy"

Last summer, Thomas Gingeras, Ph.D., a widely recognized genome investigator and developer of pioneering technologies used to probe it (notably, DNA microarrays), formed a new lab at Cold Spring Harbor Laboratory. Among other things, Gingeras returned to the campus — he had headed a lab at CSHL from the late 1970s until the mid 1980s — to carry forward his work on ENCODE. As one of the consortium's five principal investigators, he has been in prime position to ponder the significance of data generated since the project's inception. Gingeras is straightforward in conceding that his interpretation of this data is nothing short of "heresy" in the eyes of many other genome scientists.

Details of the controversies arising out of ENCODE's preliminary results — a paper setting forth pilot-stage data appeared in *Nature* last fall — are frankly abstruse. In broad terms, however, it is not difficult to appreciate why they have caused a stir. In the eyes of Gingeras, the data supports a dramatically new definition of what it means to say an element of the genome is "functional." Equally surprising, the data tend to destabilize long-held assumptions, including what it means to label a stretch of the genome a "gene."

Following completion of the human reference genome, scientists who totaled up the "acreage" devoted to its 20,000-plus protein-coding genes reported a figure ranging from 1% to 2% of the genome. It became fashionable to consider the remaining 98% to 99% "junk DNA." If a sequence did not code for protein — i.e., if it was not part of a gene — it was assumed to perform no useful function.

Perhaps the chief "heresy" to arise out of ENCODE's data is this: nearly all of the genome, far from being "junk," appears to have some kind of function. In a series of papers published from 2002 through 2007, Gingeras and ENCODE collaborators, extrapolating from an analysis of 1% of the genome, concluded that an astonishing 94% of the human genome is transcribed as RNA. The question is, What are all these RNA transcripts *doing* in cells? Are they in fact doing anything of biological importance?

That last question has a partial answer, thanks to the uncovering of a previously unknown species of RNA molecules called small RNAs. These short molecules, most of them 19 to 200 nucleotides in length, have been classified into a multiplicity of subsets, according to size and presumed function within cells (only a fraction of which are now grasped). Small RNAs have been given names that most people outside of biology will find unfamiliar, e.g., microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), short interfering RNAs (siRNAs).

Gingeras estimates that several thousand different small RNAs operate in human cells. Thanks in part to discoveries by Greg Hannon and Leemor Joshua-Tor of CSHL, we now know that specific cellular machineries "slice" and "dice" non-protein-coding RNA transcripts. The products are small RNAs that act very selectively to silence gene expression, a mechanism called RNA interference, or RNAi.

The case of small RNAs helps shed light on Gingeras's view of "function" across the genome. He contends that a very long non-coding RNA transcript, many thousands of nucleotides in extent — which orthodoxy would consider non-functional since it does not code for biologically active proteins and is not conserved by evolution — should indeed be considered a functional genome element. Its only purpose may be to present the cell's various processing machineries with copious raw material from which to excise short RNA segments. A processed small RNA, though minuscule in size relative to the non-coding sequence it was cut from, is available to regulate the expression of a specific gene. If the small RNA is considered functional, then so should the giant non-coding RNA that gave rise to it, argues Gingeras.

The genome in 'RNA space'

Gingeras wants his colleagues to think about what the genome looks like in what he calls "RNA space." In other words, from the perspective not of genes or proteins, but RNAs — the entire universe of them, protein-coding and non-coding. Using a computer, Gingeras has drawn a map that, in his words, "tells us what one of the human chromosomes looks like in RNA space." The representation of chromosome 21 looks as if it has been ripped from the notebook of a teenager playing with a compass and protractor. This abstraction of intersecting arcs bouncing off the sides of concentric circles brings to light two remarkable facts.

First, in many tissues, bits of DNA sequence associated with one gene are found inside the sequence of another gene. In some cases, a gene can be observed to "start" inside another gene; but, if its sequence is spatially scattered in this manner, how then does one define the gene? Where are "genes" located if they start or stop inside other genes, which have their own presumed start and stop points? And what to make of the latter when parts of their own sequence may also be dispersed in the space of other presumed "genic regions"?

The other unexpected insight: there are many examples on chromosome 21 of genes widely separated but whose RNA products, when meticulously traced, are shown to be mixed together. Bits of one non-coding RNA transcript are found inside another. In RNA space, in other words, a network of interaction is implied among gene products that one would otherwise have no reason to believe to be associated.

ENCODE data supports a view of the genome that "is so much more complex" than prevailing models that resistance is inevitable; "I expect people will gradually accept our data, but I also expect more arguments about the semantics of what 'functional' really means," says Gingeras. "I can only carry on with my work, and be content that the field will follow where the evidence takes us."