Mining Treasures From 'Junk DNA'

The 97% of the human genome that does not encode protein has taken a bad rap. But now this so-called junk DNA is turning out to play vital roles in normal genome function

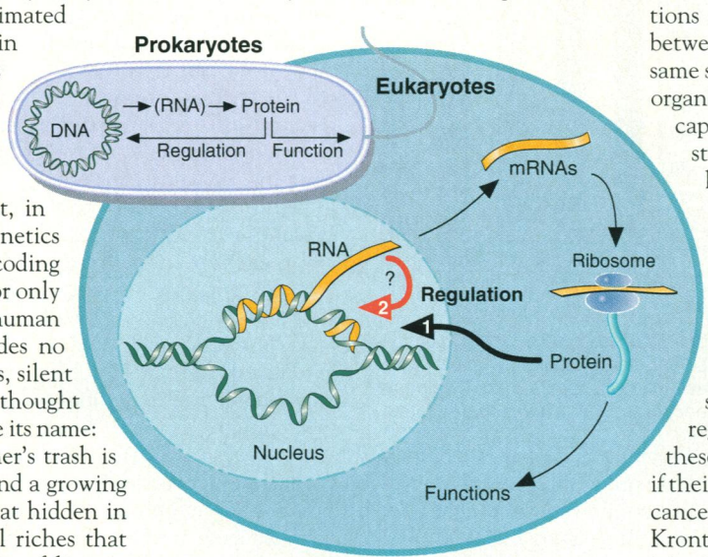
If you read the average newspaper article about recent advances in genetics, you could easily end up thinking scientists have a pretty good grip on the human genome. After all, more than 4000 genes have already been identified, and barely a week goes by without the announcement of a new addition to the repertoire. With each announcement comes the prediction that by the year 2005 every one of the estimated 100,000 human genes will be in the bag. When that's done, it's generally assumed, everything needed to understand the cell's genome, its central computer, will be known.

That's all very hopeful. But, in fact, this neat view of human genetics has a major gap. The protein-coding portions of the genes account for only about 3% of the DNA in the human genome; the other 97% encodes no proteins. Most of this enormous, silent genetic majority has long been thought to have no real function—hence its name: "junk DNA." But one researcher's trash is another researcher's treasure, and a growing number of scientists believe that hidden in the junk DNA are intellectual riches that will lead to a better understanding of diseases (possibly including cancer), normal genome repair and regulation, and perhaps even the evolution of multicellular organisms.

Rather than the genes, junk DNA "is actually the challenge right now," says Eric Lander of the Massachusetts Institute of Technology, who is himself a prominent Human Genome Project researcher. And in rising to meet that challenge, geneticists are beginning to formulate a new view of the genome. Rather than being considered a catalogue of useful genes interspersed with useless junk, each chromosome is beginning to be viewed as a complex "information organelle," replete with sophisticated maintenance and control systems—some embedded in what was thought to be mere waste.

It's not really surprising that genes have hogged the spotlight until quite recently. For one thing, the flurry of research activity that followed the cracking of the genetic code in the 1950s and '60s concentrated on simple organisms such as bacteria. And their genomes, unlike those of higher organisms, are practically wall-to-wall genes, with little junk.

Furthermore, the genes carry the genetic information for the proteins, which provide both the building blocks for cells and the enzymes that do their chemical work. Therefore, even when geneticists started studying complex, multicellular organisms, it was easy to dismiss the vast reaches of non-protein-coding DNA as a wasteland. Now, however, that notion is being overturned as researchers find that junk DNA is not a single midden heap (see sidebar), but a complex mix of different types of DNA, many of which are vital to the life of the cell.



Doubly regulated. Multicellular eukaryotes may have two systems for regulating the genome—one using proteins, the other RNAs.

ILLUSTRATION: K. SUTLIFF

Coming up in the world

Some of the earliest indications that junk DNA might have important functions came from studies on gene control. Those studies found that genes have regulatory sequences, short segments of DNA that serve as targets for the "transcription factors" that activate genes. Many of the regulatory sequences lie outside the protein-coding sequences—in the genetic garbage can. "There's at least five regulatory elements for each [human] gene, probably many more," says gene control expert Robert Tjian of the University of California, Berkeley. "For a long time it wasn't appreciated how widespread those elements can be, but now it seems that patches of really important regulatory elements can be buried among the junk DNA."

These key regulatory elements can even occur in what many geneticists have considered the ultimate in genetic detritus: the repetitive sequences scattered throughout the genomes of higher organisms. These genetic stutters have come to epitomize junk because their structures are simple to the point of absurdity, sometimes including only two or three nucleotides repeated thousands of times. In addition, the lengths and compositions of these repetitions often vary wildly between species, between organisms of the same species, even between cells of the same organism. To the average geneticist, such caprice stands in sharp contrast to the structure of crucial genes, which are known to be highly conserved in the course of evolution—precisely because their functions are so crucial. Stretches of DNA that vary so wildly, it was thought, surely cannot have an important function.

Now, however, it appears some repetitive sequences may contain stretches of DNA needed for gene regulation. What is more, the function of these stretches must be significant, because if their sequences go astray they may result in cancer. Last August, a team led by Theodore Krontiris of Tufts University School of Medicine in Boston confirmed hints that a mutation in a highly repetitive sequence called a minisatellite may contribute to as many as 10% of all cases of breast, colorectal, and bladder cancer, and acute leukemia. The minisatellite implicated in the cancers is located just after the end of the Harvey *ras* gene, which encodes a protein found in one of the cell's major growth regulatory pathways.

Krontiris says his team currently favors the idea that the minisatellite mutation pumps up *ras* gene activity, predisposing a cell to increased growth and thus to cancer. As evidence, he cites his team's findings that the mutation occurs in the part of the minisatellite that includes a regulatory sequence. "We've shown that the mutant minisatellite can bind a transcriptional regulatory factor," Krontiris says. The next step is to find out whether that binding activates the *ras* gene, as their hypothesis predicts.

But housing sequences that control the genes isn't the only role that so-called genetic trash plays. Some repetitive sequences also seem to have a crucial function in maintaining the structure of the genome. Clustered at the centers and tips (or telomeres) of

each chromosome is satellite DNA, similar to minisatellite DNA, but generally occurring in longer stretches. In November, Lisa Sandell and Virginia Zakian of the Fred Hutchinson Cancer Research Center in Seattle reported in *Cell* that when they removed telomeric DNA from yeast chromosomes, the chromosomes disintegrated. The telomeric repetitive DNA apparently protects chromosomes by binding to proteins that stop the ends of the chromosome from "fraying" and also by helping to repair damaged tips.

Introns get a boost, too

Thus, in a dramatic reversal, the repetitive sequences, once thought to be the epitome of genetic debris, now seem to be needed to maintain the integrity of the chromosomes. But the repetitive sequences aren't the only forms of genetic garbage that are moving up in the world. Whereas the repetitive sequences are usually found outside genes, a second type of genetic junk, the introns, are scattered through the genes of higher organisms. The introns, however, do not code for protein; enzymes snip them out and discard them before the coding sequences, or exons, are stitched together into the messenger RNAs that direct protein synthesis.

Introns have long been thought to be nonfunctional on the basis of an argument much like that applied to repetitive sequences: because they seem to vary willy-nilly, even between closely related species. But while that may be true for many introns, it's apparently not true for all of them. New and as yet unpublished data from Ben Koop of Canada's University of Victoria and Leroy Hood of the University of Washington in Seattle show that the introns of some genes actually show remarkable conservation between species. (Their work is in press at *Nature Genetics*.)

Koop and Hood have found that the DNA of the T cell receptor complex, a crucial immune system protein, shows 71% identity between humans and mice. That finding is startling, since only 6% of the DNA encodes the actual protein sequence, while the rest consists of introns and noncoding regions. "[The finding] certainly questions the assumption that introns are junk," says Koop. Instead, he says, "it fits the view that chromosomes are information organelles that carry out a variety of functions besides encoding genes, such as maintenance of genome structure and gene regulation."

That opinion appeals to John Mattick, a molecular biologist at the University of Queensland in Australia, currently on sabbatical at Cambridge University in England. Mattick has proposed that introns provide a previously unsuspected system for regulating gene expression. If it exists, he says, such a system could help to explain one of the key events in all evolution: metazoan radiation, the explosion of multicellular life forms that took place 530 million years ago.

The multicellular life forms have been so successful in part because they are able to coordinate far more complex sets of genes than simple prokaryotic organisms, such as bacteria. Mattick suggests that the multicellular eukaryotes (organisms with nuclei) were able to expand beyond the upper limit of 6000 to 8000 genes seen in the single-celled prokaryotes (which lack a nucleus) because of the presence of not just one, but two sets of gene regulatory systems.

In the first system, shared by both eukaryotes and prokaryotes, proteins encoded by genes modulate the activity of other genes by feeding back on regulatory sites in the DNA. But, Mattick argues, the prokaryotes never gained the advantages conferred by the second system. He proposes that the second system is encoded in the introns, and prokaryotes couldn't tolerate introns in their genomes.

Evolutionary theorists agree that introns are the remnants of primitive RNA life forms that somehow became incorporated in cells,

Talkin' Trash: A Glossary of Junk DNA

Genes are DNA sequences that either carry the information for making proteins or play some other direct role in protein synthesis (such as making the transfer RNAs that help assemble amino acids into proteins). But in higher species, the overwhelming majority of the DNA—97% in humans—does not code for proteins or RNAs with clear functions.

Many geneticists, uncertain about what to make of this apparently superfluous DNA, have taken to referring to it as "junk." But what looks like junk can hide gems, and there is new and growing evidence that some junk DNA may be of great value (see main story). Even in the junkpile, however, there are distinctions to be made. Researchers have found several different types of junk DNA and its companion RNA. Here is the main glossary.

- Introns: Most genes of higher organisms are interrupted by sequences called introns that don't code for proteins. Introns may play other vital roles, however. For instance, a slew of so-called small nucleolar RNAs (snoRNAs) are encoded by introns. Because snoRNAs accumulate in the nucleolus, where the protein-making ribosomes are formed, researchers speculate that they play a role in ribosome assembly.
- Satellites: These short DNA sequences are repeated hundreds or thousands of times at a stretch, mainly at the ends and centers of chromosomes. Although they look like the quintessence of waste matter, in fact, the survival of the chromosome may depend on satellite DNA.
- Minisatellites: Similar to satellites, but shorter, these repetitive sequences are found throughout the genome. They can be dangerous—defects in some minisatellites are associated with cancer.
- Microsatellites: Even shorter than minisatellites. Defects in microsatellites are also associated with cancer.
- 3′ Untranslated regions (3′UTRs): The final protein coding portions of genes are followed by DNA that is transcribed into RNA but is not translated into protein. The 3′UTRs won new respect when researchers discovered that they contain sequences that regulate gene activity.
- Heterogeneous nuclear RNA (hnRNA): About 25% of the hnRNAs are immature messenger RNAs, containing both the exons and introns of genes. The other 75% is still a mystery.
- Short interspersed elements (SINEs): Numerous copies of these repetitive sequences clutter the human genome. For example, the 300-base-pair "Alu" sequence (named for the enzyme that helped identify it) occurs 500,000 times. They are generally ignored as nonfunctioning DNA of parasitic origin. But SINEs hop about the genome, and, if they land in a gene, may cause disease. This kind of disruption has caused neurofibromatosis-1, or elephant man's disease.
- Long interspersed elements (LINEs): Similar to SINEs but longer—up to 7000 base-pairs apiece. Also assumed to be nonfunctional; may also cause disease through gene disruption.
- Pseudogenes: These defective copies of genes lack introns and are rarely, if ever, expressed.

–R.N.

and evolved with them. In Mattick's scheme, the prokaryotes couldn't put up with the introns, because their cells lack nuclei and, as a result, the transcription of DNA into messenger RNA and the translation of messenger RNA into protein occur simultaneously. Hence if introns were introduced into a prokaryotic cell's genes, there would be no opportunity to remove them before the protein is made, and the result would be "nonsense" nonfunctional proteins. As Mattick puts it, for prokaryotes, "the selection pressures against carrying parasitic RNA elements would be enormous."

In eukaryotic cells, on the other hand, transcription takes place within the nucleus, and translation in the surrounding cytoplasm. If introns are present in the genes, they can be snipped out before translation takes place. And once the introns had set up shop in eukaryotic genes, argues Mattick, they got co-opted into providing a second mechanism for regulating the genome—one that may explain the enormous evolutionary success of eukaryotes. This system works according to a quite different mechanism from its protein-based counterpart: In the second system Mattick is proposing, genes are regulated by intron-encoded RNAs that bind either the DNA or RNA.

"[Mattick's] idea is very interesting indeed," says evolutionary geneticist Laurence Hurst of Cambridge University, England. "And it's perfectly testable." For example, he says, Mattick's model predicts that certain genes, like regulatory developmental genes, that must be finely controlled, will likely bear intron-encoded regulatory RNAs.

In fact, surprising new results point to the existence of just such regulatory RNAs. In the fall of 1992, researchers found that the XIST gene, which shuts down one of the two X chromosomes in female mammalian cells, performs its function without ever making a protein. Now, a team led by Jeanne Lawrence of the University of Massachusetts Medical School in Worcester has evidence suggesting that XIST makes an RNA that remains stuck to the chromosome, blocking its further activity.

That intriguing finding provides support for Mattick's revolutionary idea. Late last year, Rosalind Lee, Rhonda Feinbaum, and Victor Ambros of Harvard University offered up another example of a regulatory RNA. The Harvard group found that *lin-4*, a developmental control gene from the roundworm *Caenorhabditis elegans*, encodes a small RNA that binds to the messenger RNA of another gene called *lin-14*, blocking its ability to make a protein. This discovery was particularly dramatic, given Mattick's gene regulation hypothesis, because *lin-4* sits in the intron of another gene.

That was an interesting bit of confirmation for a novel—and not yet widely accepted—theory. But it doesn't exhaust the junk DNA connections with *lin-4*. The binding site of the lin-4 RNA on the *lin-14* RNA is the so-called 3′ untranslated region (3′UTR)—a region that until very recently was dismissed as deadly dull junk. The 3′UTR, which lies at the end of each gene's messenger RNA, is not translated into protein, and for that reason it had been classified as functionless.

This "nonfunctional" region, however, seems to provide the site for some important regulatory activities, since it may not just be *lin-14*'s activity that is regulated via interactions with its 3′UTR end. Over the past few years, a swarm of discoveries has revealed that mutations in the 3′UTR region of at least 10 different genes from worms, fruit flies, and vertebrates can suppress the activities of those genes, by suppressing translation or by hastening degradation of their messenger RNAs. Indeed, one of those mutations, which occurs in the 3′UTR region of the gene for the enzyme myotonin kinase, triggers myotonic dystrophy, a hereditary muscle-wasting disease.

What's more, *lin-4*, *XIST*, and the few other regulatory RNAs that are identified may be just the tip of the iceberg. "There's too many cases of odd RNAs," says molecular geneticist Marvin Wickens of the University of Wisconsin, Madison. "It smells like there might be a whole family of regulatory RNAs." And if that suspicion proves correct, it would be a big boost for Mattick's new theory, as well as for the status of junk DNA—a status that is likely to keep on rising over the next couple of years. Enough gems have already been uncovered in the genetic midden to show that what was once thought to be waste is definitely being transmuted into scientific gold.

–Rachel Nowak

Additional Reading

J.W. Bodnar, "Telephone Book of Life," *Nature* **361**, 580 (1993).

M. Wickens and K. Takayama, "Deviants—or Emissaries," *Nature* **367**, 17.

EMBRYOLOGY

Watching New Developments—Live

In the Golden Age of television—its embryonic years—most shows were broadcast live. That often made them more exciting than today's taped shows, since the audience could delight in watching the performers respond to the inevitable surprises on the set.

Today, an anxious new group is awaiting the unexpected developments that come with live broadcasts: embryologists. On page 681, Russell Jacobs and Scott Fraser of the California Institute of Technology describe a new magnetic resonance imaging (MRI) technique that may allow researchers to watch the movements of all the cells of developing embryos—live. "They have done something that is very, very difficult and done it exquisitely. They have taken [microscopic MRI] into a discipline where it will show a lot of promise," says G. Allan Johnson of Duke University's Center for *In Vivo* Microscopy, one of the pioneering centers of microscopic MRI.

The new technique uses a souped-up MRI machine that has a magnetic field 10 times stronger than those of clinical instruments and can therefore provide a million times better resolution—enough to see even individual cells provided they are labeled with a suitable contrast agent. After the Caltech team injected such an agent into a single cell of a live, 16-cell frog embryo, they watched what happened to that cell and its progeny as the organism developed. "We can look at the whole embryo, inside and outside, and watch how sheets of cells move in relation to each other," Jacobs says.

Such cellular movements and interactions are critical for normal development of complex organisms, and even though biologists have already been remarkably successful in tracing the cellular choreography in embryos, their conclusions have been limited by the available techniques. Following the fate of cells inside an embryo usually requires the examination of many different embryos, each halted at a different stage of development, fixed, and then analyzed microscopically.

Microscopic MRI has changed that, say Jacobs and Fraser, and its newfound capacity for imaging all the cells of the embryo is bound to bring new insights. In fact, Jacobs and Fraser have already made one observation that challenges the received wisdom. They noticed that movements of two key early layers of cells, the mesoderm (which gives rise to blood and muscles) and the ectoderm (which produces skin and nerves) may not be as well coordinated as had been assumed. The promise of new insights is heightened by Jacobs and Fraser's more recent finding that the technique can even be used on mouse embryos within the uterus.

Despite its promise, don't expect microscopic MRI to become a staple in every lab. Richard Harland, a developmental biologist at the University of California, Berkeley, notes that the technology is expensive and there's skepticism about exactly what new biological questions it can address. Yet Harland says he's intrigued. "There's undoubtedly a tremendous advantage to watching it happen live. There might be a host of surprises," says Harland. Expect embryologists to be glued to their monitors in the future.

–John Travis