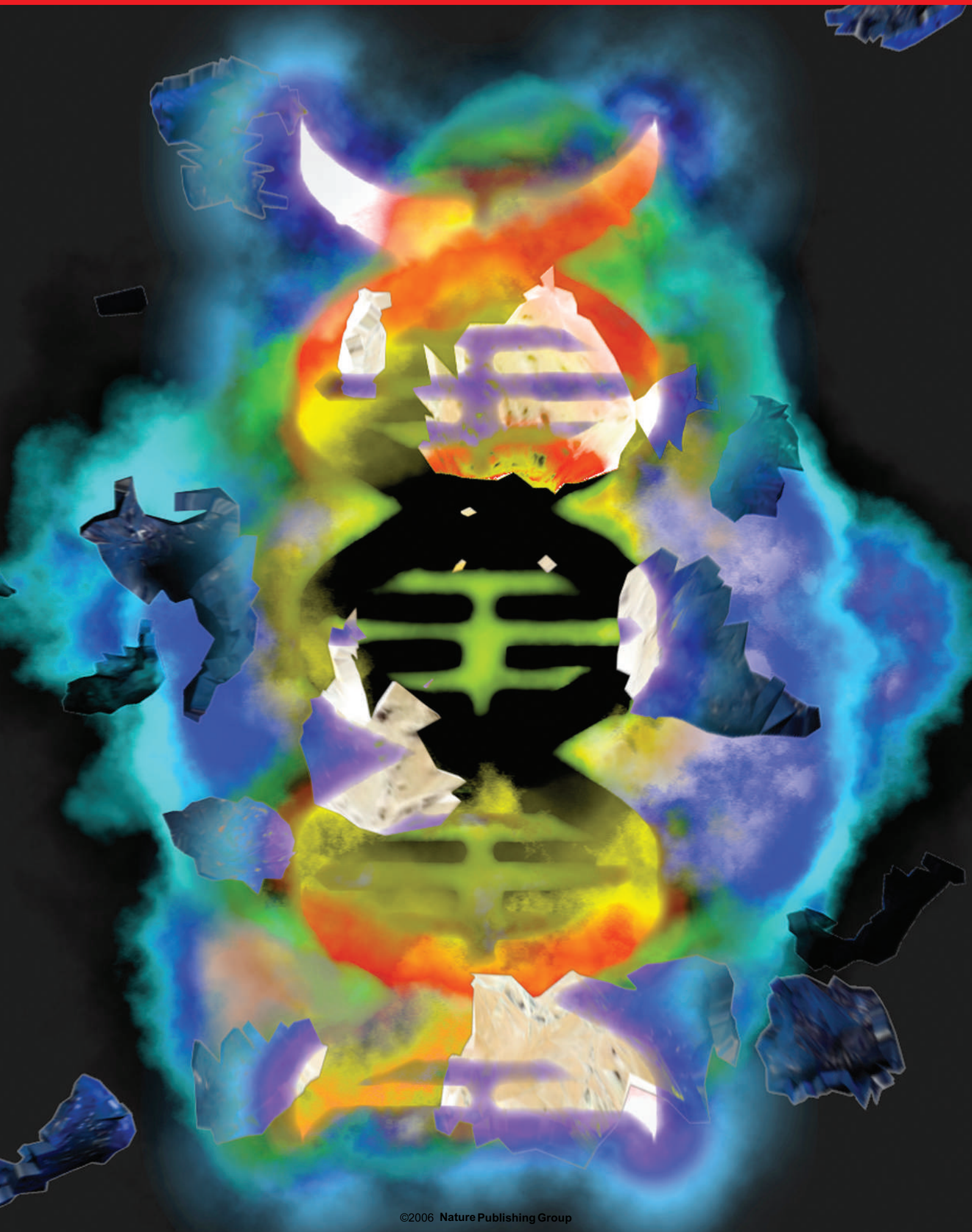
NATURE|Vol 441|25 May 2006



C. DARKIN

WHAT IS A GENE?

The idea of genes as beads on a DNA string is fast fading. Protein-coding sequences have no clear beginning or end and RNA is a key part of the information package, reports **Helen Pearson**.

'Gene' is not a typical four-letter word. It is not offensive. It is never bleeped out of TV shows. And where the meaning of most four-letter words is all too clear, that of gene is not. The more expert scientists become in molecular genetics, the less easy it is to be sure about what, if anything, a gene actually is.

Rick Young, a geneticist at the Whitehead Institute in Cambridge, Massachusetts, says that when he first started teaching as a young professor two decades ago, it took him about two hours to teach fresh-faced undergraduates what a gene was and the nuts and bolts of how it worked. Today, he and his colleagues need three months of lectures to convey the concept of the gene, and that's not because the students are any less bright. "It takes a whole semester to teach this stuff to talented graduates," Young says. "It used to be we could give a one-off definition and now it's much more complicated."

In classical genetics, a gene was an abstract concept — a unit of inheritance that ferried a characteristic from parent to child. As biochemistry came into its own, those characteristics were associated with enzymes or proteins, one for each gene. And with the advent of molecular biology, genes became real, physical things — sequences of DNA which when converted into strands of so-called messenger RNA could be used as the basis for building their associated protein piece by piece. The great coiled DNA molecules of the chromosomes were seen as long strings on which gene sequences sat like discrete beads.

This picture is still the working model for many scientists. But those at the forefront of genetic research see it as increasingly old-fashioned — a crude approximation that, at best, hides fascinating new complexities and, at worst, blinds its users to useful new paths of enquiry.

Information, it seems, is parceled out along chromosomes in a much more complex way than was originally supposed. RNA molecules are not just passive conduits through which the gene's message flows into the world but active regulators of cellular processes. In some cases, RNA may even pass information across generations — normally the sole preserve of DNA.

An eye-opening study last year raised the possibility that plants sometimes rewrite their DNA on the basis of RNA messages inherited from generations past[1]. A study on page 469 of this issue suggests that a comparable phenomenon might occur in mice, and by implication in other mammals[2]. If this type of phenomenon is indeed widespread, it "would have huge implications," says evolutionary geneticist Laurence Hurst at the University of Bath, UK.

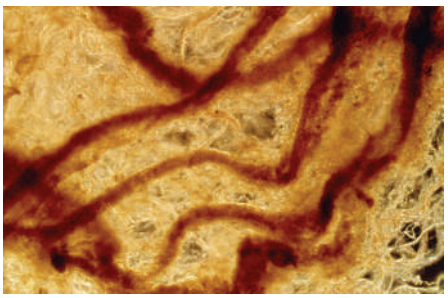
"All of that information seriously challenges our conventional definition of a gene," says molecular biologist Bing Ren at the University of California, San Diego. And the information challenge is about to get even tougher. Later this year, a glut of data will be released from the international Encyclopedia of DNA Elements (ENCODE) project. The pilot phase of ENCODE involves scrutinizing roughly 1% of the human genome in unprecedented detail; the aim is to find all the sequences that serve a useful purpose and explain what that purpose is. "When we started the ENCODE project I had a different view of what a gene was," says contributing researcher Roderic Guigo at the Center for Genomic Regulation in Barcelona. "The degree of complexity we've seen was not anticipated."

"We've come to the realization that the genome is full of overlapping transcripts."
— Phillip Kapranov

Under fire

The first of the complexities to challenge molecular biology's paradigm of a single DNA sequence encoding a single protein was alternative splicing, discovered in viruses in 1977 (see 'Hard to track', overleaf). Most of the DNA sequences describing proteins in humans have a modular arrangement in which exons, which carry the instructions for making proteins, are interspersed with non-coding introns. In alternative splicing, the cell snips out introns and sews together the exons in various different orders, creating messages that can code for different proteins. Over the years geneticists have also documented overlapping genes, genes within genes and countless other weird arrangements (see 'Muddling over genes', overleaf).

Alternative splicing, however, did not in itself require a drastic reappraisal of the notion of a gene; it just showed that some DNA sequences could describe more than one protein. Today's assault on the gene concept is more far reaching, fuelled largely by studies that show the previously unimagined scope of RNA.



Spools of DNA (above) still harbour surprises, with one protein-coding gene often overlapping the next.

P. PLAILLY/SPL

The one gene, one protein idea is coming under particular assault from researchers who are comprehensively extracting and analysing the RNA messages, or transcripts, manufactured by genomes, including the human and mouse genome. Researchers led by Thomas Gingeras at the company Affymetrix in Santa Clara, California, for example, recently studied all the transcripts from ten chromosomes across eight human cell lines and worked out precisely where on the chromosomes each of the transcripts came from[3].

The picture these studies paint is one of mind-boggling complexity. Instead of discrete genes dutifully mass-producing identical RNA transcripts, a teeming mass of transcription converts many segments of the genome into multiple RNA ribbons of differing lengths. These ribbons can be generated from both strands of DNA, rather than from just one as was conventionally thought. Some of these transcripts come from regions of DNA previously identified as holding protein-coding genes. But many do not. "It's somewhat revolutionary," says Gingeras's colleague Phillip Kapranov. "We've come to the realization that the genome is full of overlapping transcripts."

Other studies, one by Guigo's team[4], and one by geneticist Rotem Sorek[5], now at Tel Aviv University, Israel, and his colleagues, have hinted at the reasons behind the mass of transcription. The two teams investigated occasional reports that transcription can start at a DNA sequence associated with one protein and run straight through into the gene for a completely different protein, producing a fused transcript. By delving into databases of human RNA transcripts, Guigo's team estimate that 4–5% of the DNA in regions conventionally recognized as genes is transcribed in this way. Producing fused transcripts could be one way for a cell to generate a greater variety of proteins from a limited number of exons, the researchers say.

Many scientists are now starting to think that the descriptions of proteins encoded in DNA know no borders — that each sequence reaches into the next and beyond. This idea will be one of the central points to emerge from the ENCODE project when its results are published later this year.

Kapranov and others say that they have documented many examples of transcripts in which protein-coding exons from one part of the genome combine with exons from another

Hard to track


ABBEY OF ST THOMAS, BRNO, CZECH REPUBLIC

1860s After playing with pea plants, Austrian monk Gregor Mendel defines the basic rules of inheritance. Traits are determined by discrete units that are passed from one generation to the next.

1909 Danish botanist Wilhelm Johanssen coins the word 'gene' for the unit associated with an inherited trait, although the physical basis remains unknown.


GRAPHIC SCIENCE/ALAMY

1910 Thomas Morgan's work on fruitflies (right), shows that genes sit on chromosomes, leading to the idea of genes as beads on a string.

1941 George Beadle and Edward Tatum introduce the concept that one gene makes one enzyme.

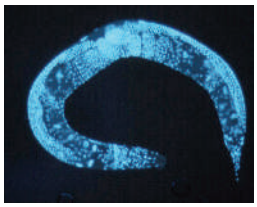
1944 Genes are made of DNA, find Oswald Avery (below), Colin MacLeod and Maclyn McCarty.


TENNESSEE STATE LIB. & ARCHIVES/NLM/NIH

1953 James Watson and Francis Crick publish the chemical structure of DNA; the central dogma of molecular biology emerges in which information flows from DNA to RNA to protein.

1977 Richard Roberts and Phillip Sharp discover that genes can be split into segments, leading to the idea that one gene can make several proteins.

1993 The first microRNA is identified in the worm *Caenorhabditis elegans*.



2003 GeneSweep: Human geneticists come up with a definition for protein-coding genes in order to decide on a winner for a bet on the number of human genes. The winner is announced, but geneticists acknowledge that they don't know the true answer.

2006 The idea that human genes are one long continuum begins to emerge.

part that can be hundreds of thousands of bases away, with several other 'genes' in between. This continuum of genes might even spill over the boundaries of chromosomes: last year, Richard Flavell at Yale University School of Medicine in New Haven, Connecticut, documented human immune-system genes that seem to be controlled by regulatory regions from another chromosome[6]. "Discrete genes are starting to vanish," Guigo says. "We have a continuum of transcripts."

Slippery concept

The large transcriptional surveys suggest that a vast amount of the RNA manufactured by the mouse and human genomes do not code for proteins. Last year a consortium of researchers in Japan, for example, estimated that a whopping 63% of the mouse genome is transcribed[7,8]; only 1–2% of the genome is thought to be spanned by sequences that contain everyday exons.

The discovery of RNA sequences that aren't just intermediates between the DNA and the protein-making machinery is not new in itself; the cell's protein-building apparatus requires a number of RNA molecules as well as proteins to operate. But the finding of 'microRNAs' and other RNA molecules now known to be vital in controlling many cellular processes in plants and animals, and the newly revealed ferment of RNA transcription, contributes to the view that RNA actively processes and carries out the instructions in the genome.

Perhaps the regions that make non-coding RNA should also carry the status of genes, if not the name itself. "I think it's time for people to take a deep breath and step back," says molecular biologist John Mattick of the University of Queensland in Brisbane, Australia. "A lot of the information in the system is being transacted by RNA."

"A lot of the information is being transacted by RNA."
— John Mattick

Although functions have been identified for several RNA molecules, the crux of the debate now is the extent to which all the extra RNA plays a part. It is conceivable that it is easier to overtranscribe and ignore the rubbish than to invest in systems that produce only what is needed. A study from last year, however, hints that at least some of the mass of RNAs is doing something useful.

Working at the Genomics Institute of the Novartis Research Foundation in San Diego, California, John Hogenesch and his co-workers systematically quenched the activity of more than 500 non-coding RNAs in human cells and found that eight were involved in cell signalling and growth[9].

But Hogenesh, and many other scientists, remain convinced that non-coding RNAs are much less important, functionally, than those that describe proteins; in the past, when scientists have searched for the genetic basis of a disease or other characteristic they have overwhelmingly found the underlying mutation to be in a protein-coding gene rather than in another region. "The preponderance of evidence suggests that protein-coding genes will hold their own when the day is over," Hogenesh says.

Some of the recent discoveries — that the human genome makes a continuum of transcripts and that cells produce masses of non-coding RNA molecules — have not posed much of a problem to people outside the world of molecular biology. Population geneticists can examine how a trait is passed down and evolves regardless of the precise molecular mechanism that underlies it. For example, geneticists can build models showing how a mutation is inherited whether it affects a protein, a non-coding RNA or a regulatory region. "I don't actually care if it's making a protein or not," says Hurst. "The equations are still the same."

But the same can't be said for studies revealing so-called extragenomic modes of inheritance. In recent years, many investigators have focused on epigenetic inheritance, in which information is passed from parent to offspring independent of the DNA sequence. And this week in *Nature* (see page 469), Minoo Rassoulzadegan's team at the French National Institute for Health and Medical Research (INSERM) in Nice, France, reports that RNA may sometimes be complicating traditional models of inheritance.

In mice, mutations in the *Kit* gene cause white patches on the tail and feet; if a mouse has one normal *Kit* gene and one mutated one it will have the spots. The odd thing is that some of the offspring of such mice, who inherit two normal *Kit* genes, still have the white tail. The French group suggest that the mutant *Kit* gene manufactures abnormal RNA molecules, which accumulate in sperm and pass into the egg. These bits of RNA somehow silence the normal *Kit* gene in the next generation and subsequent ones, producing the spotted-tail effect. "We are convinced that it's a more general phenomenon," says co-author François Cuzin.

If this is strange, the work reported last year[1] on the cress plant *Arabidopsis* by Robert Pruitt and his colleagues at Purdue University in West Lafayette, Indiana, is even stranger. Here the gene involved is called *HOTHEAD*. Pruitt and his co-workers' analysis shows that some plants do not carry the mutant version of *HOTHEAD* that their parents possessed. These plants had replaced the abnormal DNA sequence with the regular code possessed by earlier generations. "It's like, whoa, this changes everything," Pruitt says. "It definitely changes my view of inheritance."

Pruitt is now working to explain how the


M. ALICE WEBB

Back-up copies: mutant DNA in the cress plant may be 'corrected' by inherited RNA.

plant could perform such a feat. One idea is that they carry a back-up copy of their grandparents' genetic information encoded in RNA that is passed into seeds along with the regular DNA and is then used as a template to 'correct' certain genes. Conceivably, Pruitt says, some of the mystery non-coding transcripts could be responsible. "I think there's something being inherited outside what we think of as the conventional DNA genome."

Changing views

The implications of such findings for our understanding of evolution have yet to be figured out. But research into the role of RNA as a carrier of information across generations promises to enrich — and complicate — the notion of a gene yet further.

Leaving aside the can of worms that studies on epigenetics are beginning to open up, does it matter that many scientists not directly concerned with molecular mechanisms continue to think of genetics in simpler terms? Some geneticists say yes. They worry that researchers working with an oversimplistic idea of the gene could discard important results that don't fit. A medical researcher, for example, might gloss over the many different transcripts generated by a sequence at one location. And the lack of a clear idea of what a gene is might also hinder collaboration. "I find it sometimes very difficult to tell what someone means when they talk about genes because we don't share the same definition," says developmental geneticist William Gelbert of Harvard University in Cambridge, Massachusetts.

Without a clear definition of a gene, life is also difficult for bioinformaticians who want to use computer programs to spot landmark sequences in DNA that signal where one gene ends and the next begins. But reaching a consensus over the definition is virtually impossible, as Karen Eilbeck can attest. Eilbeck, who works at the University of California in Berkeley, is a coordinator of the Sequence Ontology consortium. This defines labels for landmarks within genetic-sequence databases of organisms, such as the mouse and fly, so that the databases can be more easily compared. The consortium tries, for example, to decide whether a protein-coding sequence should always include the triplet of DNA bases that mark its end.

Eilbeck says that it took 25 scientists the better part of two days to reach a definition of a gene that they could all work with. "We had several meetings that went on for hours and everyone screamed at each other," she says. The group finally settled on a loose definition that could accommodate everyone's demands. (Since you ask: "A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions.")

Rather than striving to reach a single definition — and coming to blows in the process — most geneticists are instead incorporating less ambiguous words into their vocabulary such as transcripts and exons. When it is used, the word 'gene' is frequently preceded by 'protein-coding' or another descriptor. "We almost have to add an adjective every time we use that noun," says Francis Collins, director of the National Human Genome Research Institute at the National Institutes of Health in Bethesda, Maryland.

But however much geneticists struggle to pin down the elusive gene, it is precisely its ambiguous nature that fuels their continued curiosity. "It's ever more fascinating," says Whitehead's Young. Some things, it seems, are not best portrayed by a crude four-letter word. ■

Helen Pearson is a reporter working for *Nature* in New York.

1. Lolle, S. J., Victor, J. L., Young, J. M. & Pruitt. R. E. *Nature* **434,** 505–509 (2005).
2. Rassoulzadegan, M. *et al. Nature* **441,** 469–474 (2006).
3. Cheng J. *et al. Science* **308,** 1149–1154 (2005).
4. Parra, G. *et al. Genome Res.* **16,** 37–44 (2006).
5. Akiva, P. *et al. Genome Res.* **16,** 30–36 (2006).
6. Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R. & Flavell, R. A. *Nature* **435,** 637–645 (2005).
7. FANTOM Consortium and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) *Science* **309,** 1559–1563 (2005).
8. RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium *Science* **309,** 1564–1566 (2005).
9. Willingham, A. T. *et al. Science* **309,** 1570–1573 (2005).

Muddling over genes

Science philosophers Karola Stotz, at Indiana University in Bloomington, and Paul Griffiths, now at the University of Queensland in Australia, are attempting to measure the extent of working biologists' bewilderment over genes.

They collected together 14 weird and wonderful (but real) genetic arrangements and asked biologists to decide whether each represents one, or more than one, gene.

One is a DNA segment that uses some of the same protein-coding sequences to manufacture two entirely different proteins with distinct functions. In another, one 'gene' is nestled within the non-protein coding intron of another. Another protein is assembled when four different RNA molecules, made from DNA scattered over 40,000 base pairs, are assembled into one transcript.

Confused? So were the 500 biologists who completed the questionnaire. Stotz and Griffiths found that 60% are typically sure of one answer, and 40% are confident of another. Hardly any confess that they don't know.

Stotz wants to examine whether scientists working in separate disciplines tend to view the situations in different lights. "It will be interesting to know if there is some order to the confusion," Stotz says. **H.P.**