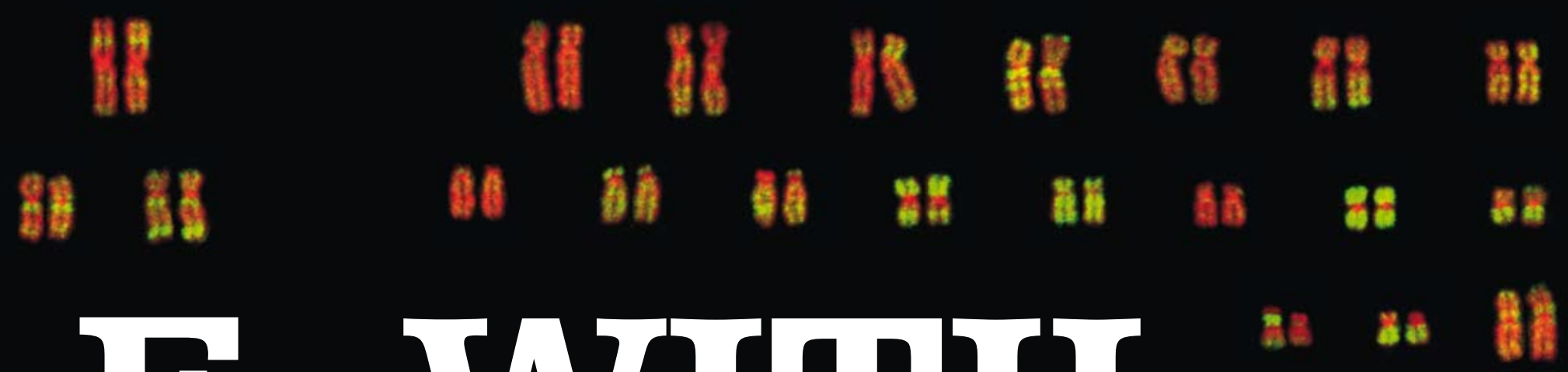


THE TROUBLE WITH GENES



The 23 pairs of human chromosomes. Highlighted in green is alu, a 'junk' DNA sequence that distinguishes primates from other mammals and now occupies 10.5% of the human genome.

Junk DNA was once thought to be little more than gibberish; evolutionary debris that puffed up our genomes. But, as **Elizabeth Finkel** reports, junk DNA may actually be the software that controls a complex organism.

WHAT'S A GENE, DAD?" I'd like to be there when the nine-year-old son of iconoclastic geneticist John Mattick pops the question. It used to be simple – a gene coded for a protein. But when I put that question to Mattick, based at the University of Queensland, his response was as disturbing as it was confusing: "Genetic information is multilayered and a gene can convey lots of different information into the system. It's almost like we've moved into hyperspace in terms of information coding and transfer."

Mattick's cutting-edge theories about gene regulation have been published in the British journal *Nature* and even appeared in the *New York Times*. Yet, even though I was once a geneticist, I couldn't fathom his answer. It seemed my fears had been realised and I'd been left behind by the genetics revolution.

In a desperate ploy to catch up, I asked how he would explain a gene to his young son. He retorted: "I would just tell him, 'it's an old-fashioned concept'; and then explain about

information networks. He's a child of the digital generation – he won't have any trouble with it."

It's not just me who's confused. I checked the 2008 edition of my favourite text book, *Molecular Biology of the Cell*. The traditional definition is still there in the opening chapter. But as you read on, you sense the textbook struggling, trying to wrestle the gene back into the box of a definition. Mattick prefers not to try. And a lot of other geneticists are starting to think this way too. As Ed Weiss at the University of Pennsylvania told me, "the concept of a gene is shredding".

Scientists were shocked when they found out how few 'old-fashioned' genes we actually have – about the same number as the humble nematode worm.

THE GENOMICS REVOLUTION is largely to blame. Scientists were shocked when they found out how few 'old-fashioned' genes we actually have – about the same number as the humble nematode worm (*Caenorhabditis elegans*). In fact, almost all multicellular creatures with the complexity of a worm or greater have about 20,000 genes. But for Mattick, the death knell of the traditional concept of the gene was triggered by another revolution altogether – that of the digital information age.

PHOTOLIBRARY, UNIVERSITY OF QUEENSLAND

Scientists have always understood biology in terms of the technology of the day. The brain, for instance, was considered by the Ancient Greeks and Romans to be an aqueduct for pumping blood; inhabitants of the 19th century likened it to a telephone exchange; those of the 20th century likened it a personal computer. Now scientists compare the brain to chaos and distributed functions of the Internet.

When it comes to the gene, Mattick likes to point out that scientists cracked its code in the 1950s, when the world was purely analogue. We had vinyl records, slide rules and mechanical cars. We were primed to recognise the gene as a recipe for an analogue device – such as a protein, for instance. Proteins are the analogue devices that operate the chemistry of life: the enzymes that metabolise food; the mortar and bricks of tissues; >>



Humans have about the same number of 'old-fashioned' genes as the nematode worm, *C. elegans* – used by biologists to study genes and development.



John Mattick, a maverick geneticist, addresses an audience of first year University of Queensland biology students.



>> the motors of muscles; the hormones that transmit signals; and the ferries that carry oxygen through blood. We recognised a gene as being the recipe for a protein.

Today, iPods store the equivalent of many thousands of vinyl records. Microprocessors in cars can control everything from the engine to the stereo. The digital revolution has succeeded in taking vast amounts of information and compressing it. Mattick believes something very similar happened to the gene. In the course of evolution, it went digital.

In 1953 we got our first inkling of how genes work. Scientists knew that genetic information was carried by the threadlike molecule DNA – a polymer of four repeating molecules adenine, thymine, cytosine and guanine or A, T, C and G. But how did this thread carry genetic information? Perhaps a picture would reveal its secret.

British crystallographers Rosalind Franklin and Maurice Wilkins bombarded crystals of DNA with X-rays and observed an enigmatic regular structure. The University of Cambridge's James Watson and Francis Crick figured out what it was. Like the elegant spiral staircase of the Louvre in Paris, it was a double helix. And the moment they figured out the structure, the secret of life was revealed. Life copied itself by splitting the helical ladder down the middle. Each half then became a template for generating a new copy because each DNA letter on the split rung specified what its partner must be: A only paired with T; C would link only with G.

Watson and Crick had figured out how the code of life copies itself. Some five years later, Marshall Nirenberg, Har Gobind Khorana and Robert Holley in the U.S. figured out what the code means. The letters of DNA spelt words that coded for amino acids – the building blocks of proteins. Until then scientists



Watson and Crick discovered that, like the spiral staircase of the Louvre Museum in Paris, DNA was a double helix that could split down the middle and become a template for a new molecule of DNA.

had been kids pulling Tinkertoys apart, but now they had the instructions for assembling them. Mankind had discovered the awe-inspiring logic of life. Genes were made up of a string of DNA, and DNA coded for proteins. DNA happened to have a go-between, a disposable working copy called 'messenger RNA'. RNA was chemically similar to DNA, but flimsier. Just as an architect will run off copies of a blueprint, so messenger RNA was the working copy used on the protein construction site.

HAVING CRACKED THE SECRET of life, these scientists now started calling themselves molecular biologists (biologists who studied living molecules). And they became rather sanguine, so sanguine they started talking about dogmas. "We had two central dogmas that were regarded as universal truths in the '60s," said geneticist Bob Williamson, now an emeritus professor at the University of Melbourne. "The first was 'DNA made RNA made protein'. The second was that the genetic code was universal: what was true for *E. coli* would be true for an elephant".

The shock to the system came in 1977. Researchers by now were quite *au fait* with genetic code. Thanks to its universality, they could insert the predicted DNA code for a human gene into a bacterium and out would pop the correct protein. Yet no-one had ever glimpsed the 'mother code' of a human gene. It was packaged in a chromosome within the dark nucleus of the cell, like a hallowed tome in the crypt of the Vatican library.

In 1977, researchers decided to fish out the mother code for the gene that makes globin (a component of haemoglobin). But no-one was prepared for its size – the globin gene

was way larger than it ought to have been. Williamson, whose group at St Mary's Hospital in London were the first to put the human globin gene into bacteria, remarked in a *Nature* editorial: "Once again we are surprised".

The explanation was bizarre. The mother gene did indeed carry the predicted code for globin, but it was strangely interspersed with gibberish.

For instance, imagine that the predicted DNA code for globin was written with the English letters: G-L-O-B-I-N. The mother code appeared as: G-L-z-z-z-q-q-O-B-s-r-m-b-I-N.

Researchers panicked. What was this gibberish? Was the genetic code not universal after all?

But the panic soon subsided. Whatever gibberish had infiltrated the mother code, it disappeared from the working copy – the messenger RNA – by the time it got to the factory floor. Like an edited home video, the internal junk had been clipped out and the good bits spliced back together again. Indeed the process was dubbed 'splicing'. The bits that were spliced together were named 'exons'; the internal junk, 'introns'. With everything neatly

Researchers panicked. What was this gibberish? Was the genetic code not universal after all?

named and explained, "the world collectively breathed a sigh of relief," says Mattick. The hallowed central dogma had been saved.

There were lots of justifications for dismissing junk DNA as 'junk'. Not only did it lack code words for amino acids; it turned out 50% of the junk was comprised of inane repetition. These repetitious tracts seemed meaningless. But researchers had

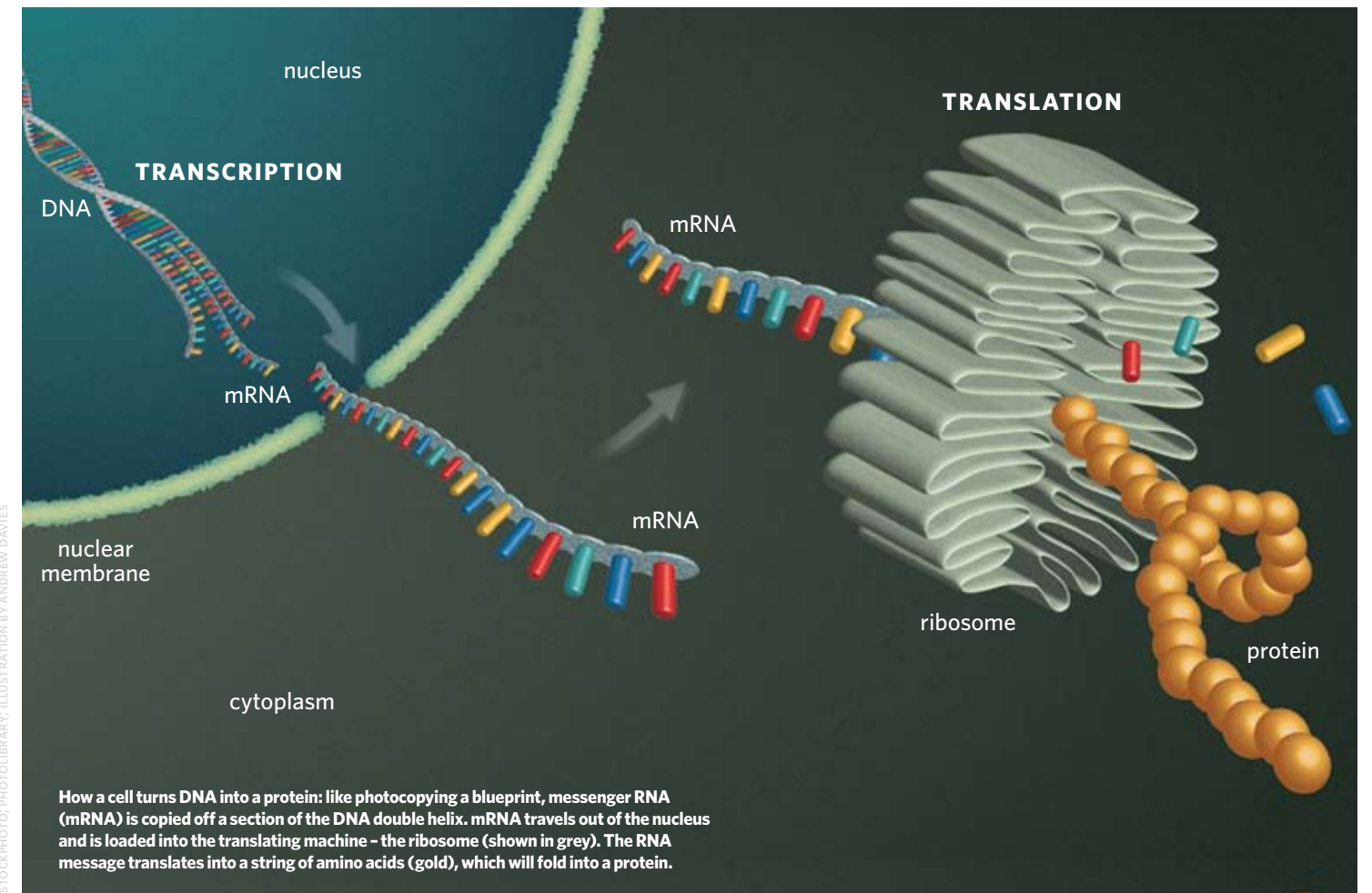
a good notion of what many of them were. Most of the repeats were 'transposons' or 'jumping genes'. Jumping genes, which may have originated from invading viruses, have the ability to copy themselves

independently of the rest of the genome and then become inserted randomly throughout the genome.

Then there was another reason to suspect that much of the DNA of a species was junk. The total amount of DNA seemed to bear very little relationship to the complexity of the organism. An amoeba for instance, had a thousand times more DNA than a human. Sometimes it seemed cells multiplied, but forgot to divide, ending up with vast amounts of DNA. It seemed as though DNA just liked to go along for the ride. >>



In the 1960s, the scientists who studied DNA and proteins began to call themselves molecular biologists. They found that the genetic code was universal - the same molecules that make up the DNA of *E. coli* make up the DNA of an elephant.



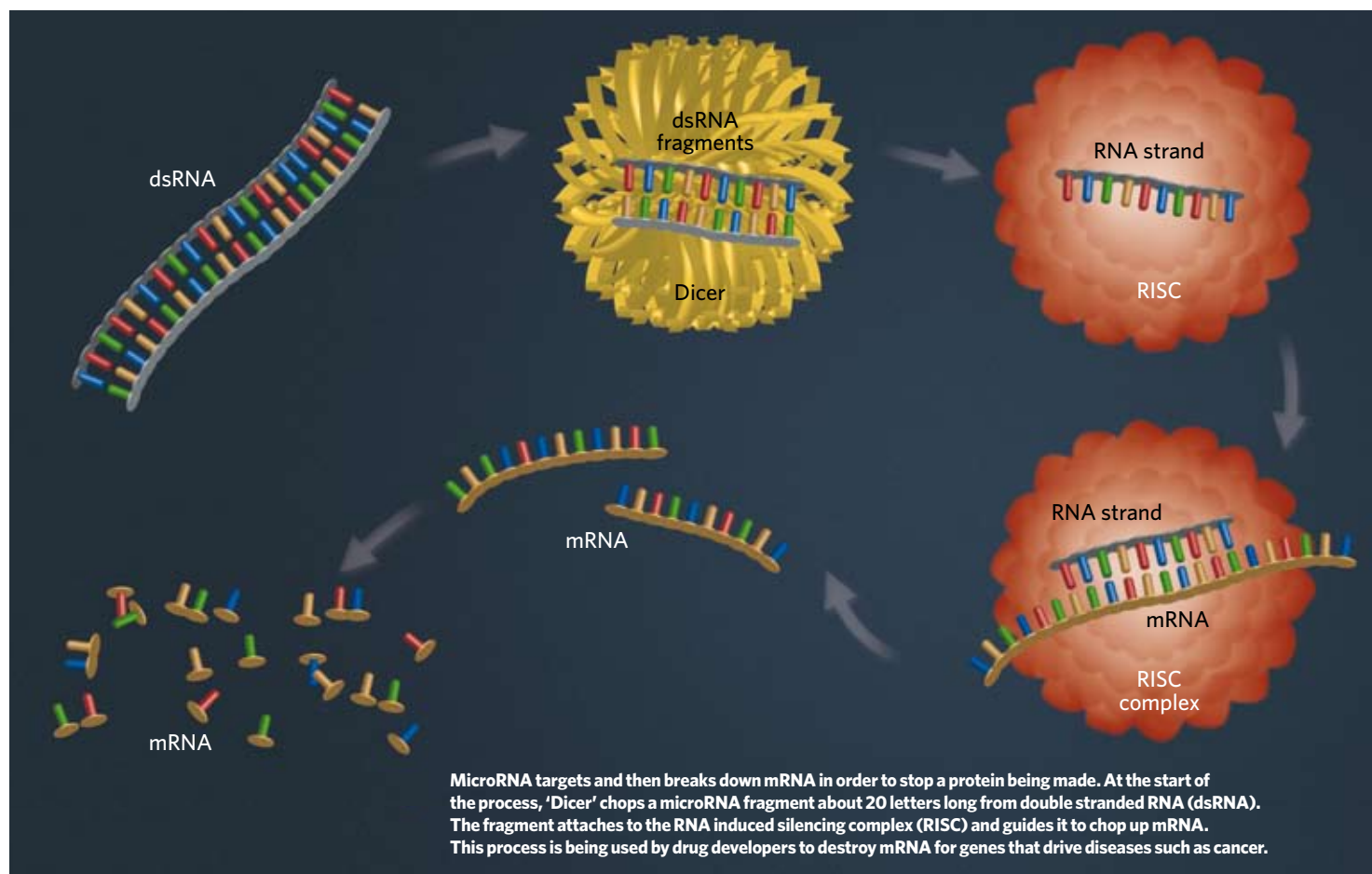


ILLUSTRATION BY ANDREW DAVIES

>> **NOT EVERYONE** dismissed junk DNA. Physicists such as Eugene Stanley at Boston University looked for patterns in junk DNA and found long-range interactions more typical of language than gibberish. Malcolm Simons, a Melbourne immunologist, stumbled upon junk DNA in the course of testing people's tissue types. Tissue compatibility depends on MHC genes, as do some aspects of immunity. Yet he found the pattern of junk DNA surrounding the genes was a better predictor of the tissue type. For him, junk turned to treasure.

There was no link between complexity and the total amount of DNA, but there was a relationship between the proportions of junk and protein-coding DNA.

Mattick's departure from the dogma seems to have been driven more by instinct than evidence. Blame it on his genes: "I've got a natural tendency to challenge everything because of my Irish background," he says. Mattick recalls sitting in a pub in 1977 during his postdoctoral stint at the Baylor College of Medicine in Houston, Texas, and thinking "Maybe this is telling us something?" But for 16 years, while he built his career as a bacterial geneticist, the problem of junk remained an "intellectual hobby".

In 1993, Mattick felt he deserved a break. He'd completed the Herculean task of setting up an entire new institute from scratch – the Institute of Molecular Biosciences at the University of Queensland in Brisbane. What better reward than to spend a sabbatical at the University of Cambridge scratching his intellectual itch?

He had slowly been building a theory in which RNA was central. The current dogma said that most of the RNA made by the genome, the RNA from introns, was bound for the scrap

When scientists found out that the seemingly simple amoeba (right) had 1,000 times more DNA than humans, many assumed the explanation must be junk DNA.



ISTOCKPHOTO: PHOTOLIBRARY

heap. But Mattick thought otherwise. Simple organisms such as bacteria do not carry introns, but complex creatures do. Mattick wondered if the scrap RNA was part and parcel of that complexity. After all, RNA has amazing versatility: it is a code-carrying molecule that can recognise matching codes on both DNA and other bits of RNA. And it can also form extraordinary three-dimensional structures to mesh with proteins.

In Mattick's theory, the scrap RNA or 'non-coding' RNA as it became called, was not flotsam and jetsam floating off a sea of junk DNA. Rather this scrap was more akin to the optical fibres of a modern high-rise building. An 18th century time traveller, spying these cables, might pass them off as scrap compared to the recognisable analogue components of the building like bathrooms, kitchens and bedrooms.

Some species of pufferfish (*Diodon*) have unusually small genomes.



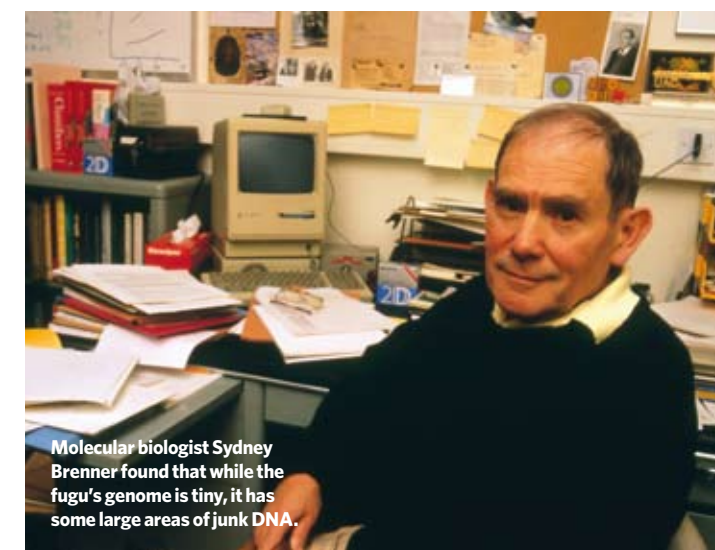
Yet, just as the cables are crucial for the building's communications and controls, so scrap RNA was crucial to the communications and control of a multicellular organism.

The major problem with his theory was that there was no experiment to prove it right or wrong. So Mattick decided to spend his sabbatical in the library looking for "circumstantial evidence". What he searched for with the most alacrity was evidence to prove him wrong. "The critical observations were the ones that would show it was bunk. Then I could just return to my lab and forget about all this stuff".

Two bits of evidence threatened to abruptly end to his quest. One was fugu – the pufferfish (*Diodon*). Fugu is famous for the tetrodotoxin which kills dozens of Japanese diners each year and for its tiny genome – about an eighth the size of our own. Nobel Prize winner Sydney Brenner, then at the British Medical Research Council's Laboratory of Molecular Biology in Cambridge, was in the process of reading fugu's DNA sequence. Rumour was the fish had barely any introns, and if a complex vertebrate such as fugu had no introns, then Mattick's theory about regulatory RNA must be wrong. He paid a visit to Brenner to discover the terrible truth. It turned out that while most of fugu's introns were very small, some were really big. Mattick's theory survived.



Mattick took the presence of non-coding RNA in a crucial stretch of the fruit fly's DNA as evidence that it plays an important role in their body formation.



Molecular biologist Sydney Brenner found that while the fugu's genome is tiny, it has some large areas of junk DNA.

The next mortal threat was a publication reporting that introns, once clipped out of the messenger RNA, were destroyed within seconds. If introns were as ephemeral as a puff of smoke, how could they perform any function? Mattick scrutinised the report closely. It showed that introns were edited out of the main message within seconds. But as to how long they persisted before being shredded, no-one knew. Perhaps, he speculated, it was long enough to do something.

Mattick, of course, was also on the lookout for evidence that would support his theory. He found some. The fruit fly possessed a set of genes that were responsible for its body plan, known as the bithorax complex. It turned out that a crucial stretch of this DNA produced RNA that did not code for protein. What other function might this RNA have?

MATTICK RETURNED TO the University of Queensland with his theory intact. He started writing papers articulating his theory that non-coding RNA (shorthand for non-protein coding RNA) was the high-level coding language of complex organisms. His approach remained one of gathering circumstantial evidence.

Together with co-workers in mathematics and computer science, he amassed some compelling observations. For instance, as more and more species became the darlings of DNA sequencing projects, Mattick noticed a delectable relationship: there was no link between the complexity of

>>

>> the critter and its total amount of DNA, but there was a clear relationship between the proportions of junk and protein-coding DNA: as the complexity of the organism increased, so did the relative amount of junk.

And then genome sequencing delivered the *pièce de résistance*: making a human being required no more old-fashioned genes than making a worm or fly. Clearly complexity was encoded elsewhere, and according to Mattick and a growing number of converts, it was in non-coding RNA.

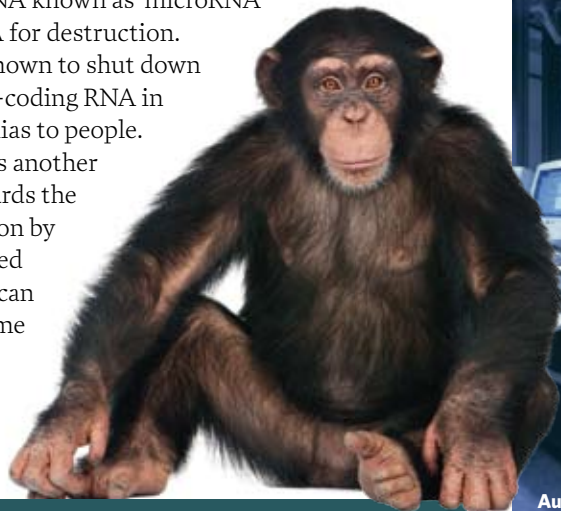
Mattick's genetic programming theory, outlined in a recent edition of the *Annals of the New York Academy of Science*, started to assume its current form. In simple terms it goes like this: bacteria could make do with using analogue devices – proteins. But even these single-celled critters devoted a large portion of their genetic information to the task of control. If organisms were going to get more complex and coordinate decisions between trillions of cells, they needed to develop a more compact regulatory language.

Just as engineers turned to digital coding to move from LPs to iPods, biological systems turned to RNA to evolve from bacteria to people. According to Mattick, RNA, like DNA, carries coded digital information in four letters that can rapidly interact with other parts of the code, much like the self-modifying or feed-forward routines of some computer programs.

AS MATTICK WAS BUILDING the framework of his model, the rest of the world started providing the bricks and mortar. Big time. Since 1993, there has been an avalanche of evidence on the surprising roles of non-coding RNA.

Most of our DNA may well have originated as 'junk' but that junk has been put to work. One of its most common jobs is to produce tiny bits of RNA known as 'microRNA' that targets other RNA for destruction. MicroRNA has been shown to shut down the activity of protein-coding RNA in everything from petunias to people.

Junk DNA also plays another crucial function: it guards the DNA code from invasion by retroviruses or so-called jumping genes, which can hop about in the genome causing dangerous mutations. Junk DNA is itself largely composed of former



Human DNA differs from that of the chimp by just 1%, and part of this divergent region is non-coding RNA that is highly active in the brain.



Mattick discovered that there is little relationship between the complexity of an organism and its amount of DNA. Humans don't require more genes than a fruit fly. But there appears to be a link between the amount of junk DNA and complexity.

interlopers but, like a patriotic immigrant, it does its best to prevent any further invasions. The RNA transcripts that run off junk DNA are still a close match to live viruses or active jumping genes, and if these junk transcripts meet up with their relatives, they inactivate them.

Junk DNA may play an even more profound role in the workings of multicellular animals. A crucial part of being multicellular is that different cells do different things – they are not all reading from the same page of the genome hymn book. The first step toward specialisation is folding down those pages that are not to be read, and it seems junk DNA guides the folding process. For instance, females carry two X chromosomes, but only read the contents of one. During embryonic development, one of the X chromosomes is folded away, a process initiated by a large string of non-coding RNA called 'Xist'.

While most tissues of the body want to keep jumping genes from jumping, the brain might have other ideas. Fred Gage's lab at the Salk Institute for Biological Studies in La Jolla, California, found evidence that jumping genes known as



Automated DNA sequencers used for the Human Genome Project took 14 years to complete the task. 'Next generation' sequencers can do the same job in hours.

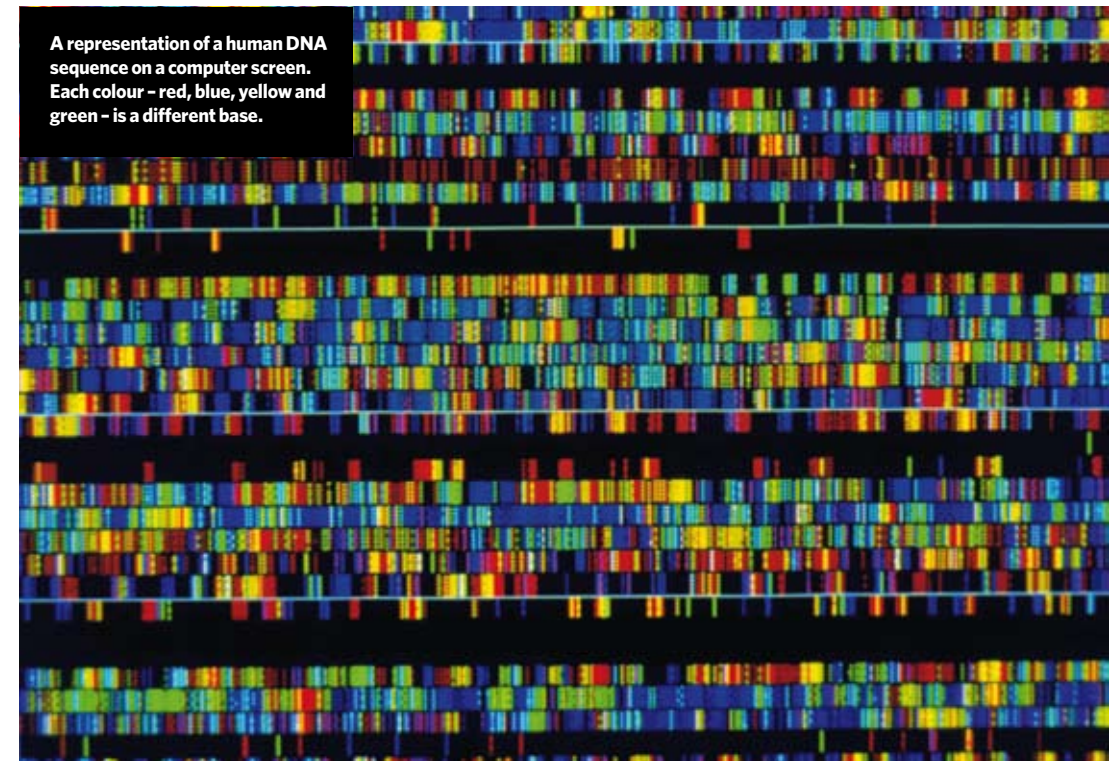
'LINE-1' or 'L1', which are permanently deactivated in other cells, become active during development of the human brain. The L1 genes replicate and insert randomly, sometimes creating as many as 100 extra copies per cell. This variation among neurons in our brains could be the basis for individual differences in neural circuitry and may open up a new way of looking at neurological disorders.

Junk may also have played a crucial role in our evolution. At the DNA level, one of the things that distinguishes primates from other mammals is the invasion of a million copies of a jumping gene that goes by the name of 'alu'. It now occupies 10.5% of the human genome. Junk RNA may also account for some of the difference between humans and chimps. Our DNA is 99% similar, but one of the regions that differs is the so-called 'HAR1', or human accelerated region 1. It turns out HAR1 produces a 118-letter non-coding RNA, which is highly active in the brain.

Junk DNA may have played a crucial role in making us human. One of the things that distinguishes primates from other mammals is the invasion of a million copies of a jumping gene.

IN 2005, MATTICK resigned as director of his institute and went back to work in the lab. Tools to explore the function of non-coding RNA had arrived in the form of heavy-duty sequencing machines. In just one week a 'next generation' sequencing machine can read three billion letters – the equivalent of an entire human genome. Not long ago, that task took the combined forces of the Human Genome Project 14 years to complete.

Mattick and University of Queensland colleague Sean Grimmond have been in collaboration with like-minds at Japan's RIKEN institute. They have been scouring the output of mouse and human genomes, trying to put together a comprehensive catalogue of their RNA output. The database called 'Fantom' (functional annotation of mammalian genomes) now contains millions of transcripts. The latest data is mind boggling. As Grimmond tells me: "Each gene is capable of seven different transcripts, some of these code for proteins



A representation of a human DNA sequence on a computer screen. Each colour – red, blue, yellow and green – is a different base.

and some don't." Trying to make sense of this deluge is the challenge. "[But] we're getting good at asking questions about ludicrous amounts of data," he says.

For Mattick, the human genome is an RNA machine. But is his theory well and truly vindicated? Not yet. Though it would be hard to find anyone today who blithely dismisses junk DNA, few are willing to go as far as he is and say that the RNA read from junk code is the software that controls a complex organism. For example,

Claude Desplan at New York University has studied fruit fly development for 25 years and argues that complex genomes, in flies or people, are still fundamentally controlled by proteins. While acknowledging that some junk has a role he says, "most of junk DNA is still junk".

Mattick, though, is convinced that our genome is way ahead of anything that IT designers have yet imagined. "The genome is so sophisticated, that there are lessons in information storage and transmission that will be really useful," once we figure it out, he tells me. "The human genome is a similar size to Microsoft Word, but it makes a human that walks and talks."

Notwithstanding the deluge of papers he has authored in top journals, Mattick still seems to be on the fringe. And you get the impression that's just where he likes it. ❧

Elizabeth Finkel is a contributing editor of *Cosmos* based in Melbourne.