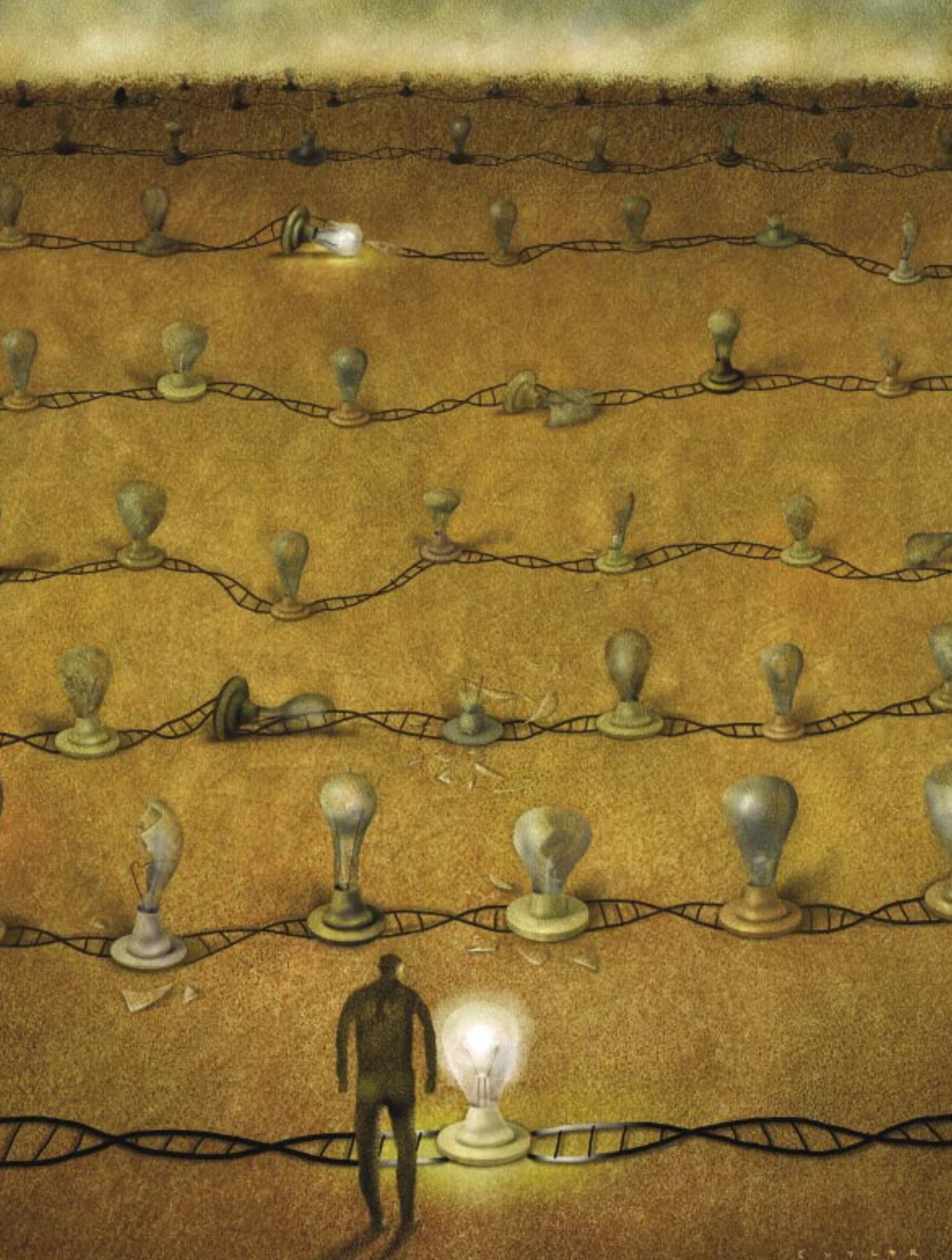


Disabled genes, molecular relics scattered across the human genomic landscape, have a story of their own to tell. And it is still unfolding

The Real Life of Pseudogenes

By Mark Gerstein and Deyou Zheng

Our genetic closet holds skeletons. The bones of long-dead genes—known as pseudogenes—litter our chromosomes. But like other fossils, they illuminate the evolutionary history of today's more familiar forms, and emerging evidence indicates that a few of these DNA dinosaurs may not be quite so dead after all. Signs of activity among pseudogenes are another reminder that although the project to sequence the human genome (the complete set of genetic information in the nuclei of our cells) was officially finished, scientists are still just beginning to unravel its complexities.

It is already clear that a whole genome is less like a static library of information than an active computer operating system for a living thing. Pseudogenes may analogously be vestiges of old code associated with defunct routines, but they also constitute a fascinating record contained within the overall program of how it has grown and diversified over time. As products of the processes by which genomes remodel and update themselves, pseudogenes are providing new insights into those dynamics, as well as hints about their own, possibly ongoing, role in our genome.

DAVE CUTLER STUDIO

Copied, Not Fake

"FALSE" GENES, which look like real genes but have no apparent function, were first recognized and dubbed pseudogenes during the late 1970s, when early gene hunters began trying to pinpoint the chromosomal locations associated with production of important molecules. For example, while seeking the gene responsible for making betaglobin, a key component of the hemoglobin protein that transports oxygen through the bloodstream, scientists identified a DNA sequence that looked like a globin gene but could not possibly give rise to a protein. Essential functional parts of the gene's anatomy were disabled by mutations, making it impossible for cellular machinery to translate the gene into a useful molecule.

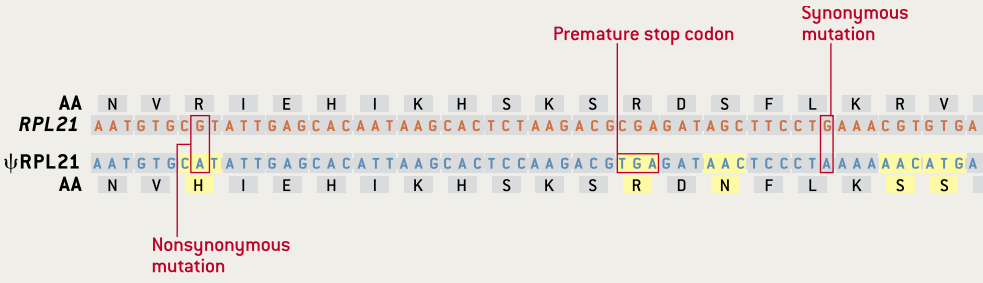
Only the far more recent completion of sequencing projects covering the full genomes of humans and other organisms allowed geneticists to get an aerial view of the genomic landscape and to appreciate how riddled with such oddities it is. The human genome is made up of more than three billion pairs of nucleotides, the building blocks of DNA molecules. Yet less than 2 percent of our genomic DNA directly encodes proteins. Perhaps a third is noncoding sequences within genes, called introns. The remaining tracts between genes constitute the great majority of our DNA, and much of it is effectively genomic dark matter whose function is still largely a mystery. It is in these seemingly barren expanses that most pseudogenes are randomly scattered like rusted car parts on the landscape—and in surprising numbers.

PSEUDOGENE BIRTH AND GENE DEATH

Two distinct processes can duplicate genes, and together they allow genomes to grow and diversify over evolutionary time. If errors in a copy destroy its ability to function as a gene, however, it becomes a pseudogene instead (*right*). The mutations that can kill a gene (*below*) range from gross deletions (such as the loss of the promoter region that signals the start of a gene sequence) to minute changes in the DNA sequence that skew the meaning of the gene's protein-encoding segments, called exons.

GENE DEATH

Genes die and become pseudogenes when mutations generated during the gene-copying process or accumulated over time render them incapable of giving rise to a protein. Cellular machinery reads the DNA alphabet of nucleotide bases (abbreviated A, C, G, T) in three-base increments called codons, which name an amino acid building block in a protein sequence or encode "stop" signals indicating the end of a gene. Even single-base mutations in codons can alter their amino acid meaning, and base deletions or insertions can affect neighboring codons by shifting the cellular machinery's reading frame. The alignment shown here of a partial sequence for a human gene (*RPL21*) against one of its pseudogene copies (ΨRPL21), along with each codon's corresponding amino acid (AA), illustrates some of the disabling mutations typically found in pseudogenes.



With ongoing annotation of the human genome sequence, our research group, along with others in Europe and Japan, have identified more than 19,000 pseudogenes, and more are likely to be discovered. Humans have only an estimated 21,000 protein-coding genes, so pseudogenes could one day be found to outnumber their functional counterparts. Their sheer prevalence has raised many questions, including how they came into existence, why there are so many of them and why, if they are really useless, they have been retained in our genome for so long.

The answer to the first question is already fairly well understood. A small fraction of pseudogenes are believed to have once been functional genes that simply "died" from disabling changes to their nucleotide sequences. But most pseudogenes are disabled duplicates of working genes. They may have been dead on arrival, having suffered lethal damage during the copying process, or they may have accumulated debilitating mutations over time that collectively rendered them incapable of functioning.

Critical to a working gene is an intact anatomy that includes uninterrupted nucleotide spans called exons, which correspond to amino acid sequences in the encoded protein. Introns typically separate the exons, and at the beginning of a gene is a segment known as a promoter that serves as the starting point for cellular machinery to recognize the

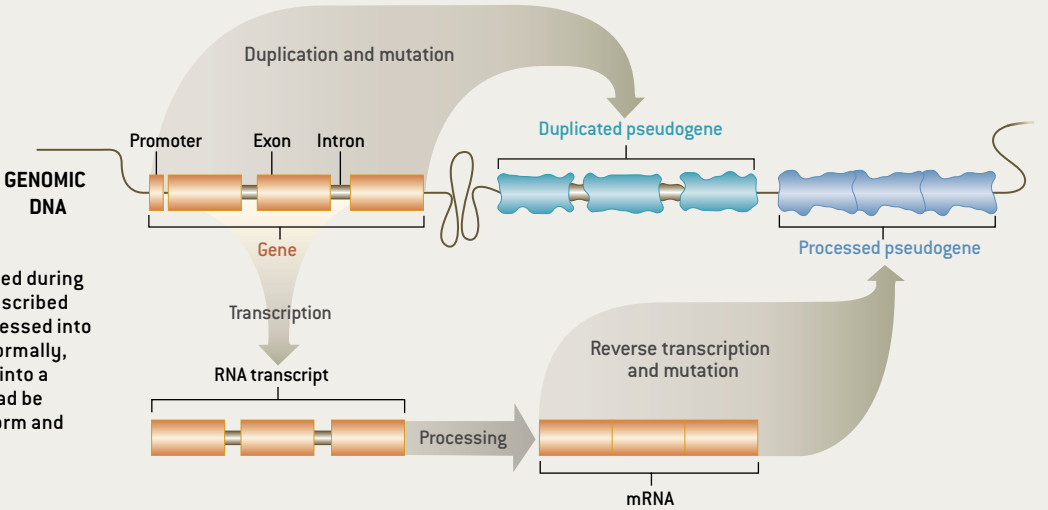
Overview/*The Pseudogenome*

- Pseudogenes are the molecular remains of broken genes, which are unable to function because of lethal injury to their structures.
- The great majority of pseudogenes are damaged copies of working genes and serve as genetic fossils that offer insight into gene evolution and genome dynamics.
- Identifying pseudogenes involves intensive data mining to locate genelike sequences and analysis to establish whether they function.
- Recent evidence of activity among pseudogenes, and their potential resurrection, suggests some are not entirely dead after all.

FLAWED COPIES

A "DUPLICATED" PSEUDOGENE arises when a cell is replicating its own DNA and inserts an extra copy of a gene into the genome in a new location.

A "PROCESSED" PSEUDOGENE is formed during gene expression, when a gene is transcribed into RNA, then that transcript is processed into a shorter messenger RNA (mRNA). Normally, the mRNA is destined for translation into a protein—but sometimes it can instead be reverse-transcribed back into DNA form and inserted in the genome.



gene on a chromosome. When a cell expresses a gene, it first recruits essential molecular players to the promoter site, which travel down the gene's length, transcribing it into a preliminary RNA copy. A splicing process next cuts introns out of the raw transcript and joins exonic sequences to produce an edited messenger RNA (mRNA) version of the gene. The mRNA is then read by a ribosome, a cellular machine that translates its sequence into the string of amino acids that forms a protein, the molecule that ultimately carries out the gene's function.

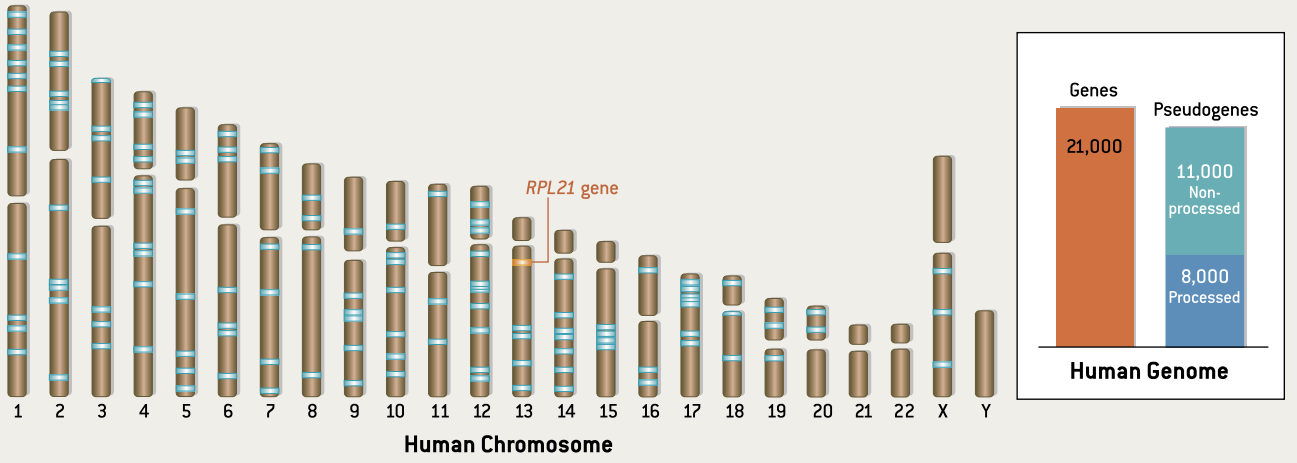
Pseudogenes can be born in two ways, each of which yields a distinctive facsimile of the original parent gene. Just before dividing, a cell duplicates its entire genome, and during that process, an extra copy of a gene can be inserted into the chromosomes in a new location. Alternatively, a new version of a gene can also be created through reverse transcription: during gene expression, the mRNA is copied back into a sequence of DNA that is inserted into the genome. Known as retrotransposition, this phenomenon can occur because of the activity of another type of transposable genetic actor, known as a long interspersed nuclear element, or LINE, that behaves like a genomic virus. LINEs carry their own machinery for making DNA copies of themselves to insert into the genome, and mRNA transcripts that are in the vicinity when LINEs are active can be swept up and retrotransposed as well.

These two processes, duplication and retrotransposition, are major forces that remodel genomes over the course of evolutionary time, generating new variation in organisms. They are the means by which genomes grow and diversify, because many replicated genes remain active. But if the gene copy contains disabling typos or is missing pieces of the original, such as the promoter, it will become a pseudogene. Those arising from duplication of an entire gene are recognizable because they contain both introns and exons. Pseudogenes made from mRNA lack introns and are described as processed pseudogenes.

Although the overall distribution of most pseudogenes across human chromosomes seems completely random, certain kinds of genes are more likely to give rise to pseudogenes. Geneticists organize functional genes into families based on their similarity to one another in both sequence and purpose. Only about a quarter of these family groups are associated with a pseudogene, and some families have spun off an inordinate number of copies. For example, the family of 80 human genes that produce ribosomal proteins has given rise to about 2,000 processed pseudogenes—roughly a tenth of the genome's identified total. In one extreme case, a single ribosomal protein gene known as *RPL21* has spawned more than 140 pseudogene copies.

This disparity probably derives from

LUCY READING-IKKANDA



PSEUDOGENE DESCENDANTS (*blue*) of the ribosomal protein gene *RPL21* (*orange*) are scattered across the human chromosomal landscape. Overall distribution of pseudogenes in the human genome appears to be completely random, although some local genome regions tend to contain more pseudogenes. Those DNA regions may be analogous to certain geochemical environments that better preserve mineral fossils. Identification of genes and pseudogenes is an ongoing process, but to date more than 19,000 pseudogenes have been identified in the human genome—only slightly less than the current tally of around 21,000 human genes (*inset*). About 8,000 of our pseudogenes are processed; the rest include include duplicated pseudogenes and other nonprocessed subcategories.

the activity levels of different genes. Those responsible for basic cellular housekeeping functions, such as the genes in the ribosomal protein family, are abundantly expressed, providing more opportunities to create processed pseudogenes.

Because pseudogenes have been accreting this way in our genomes for so long, some are relics of genes eliminated during the course of evolution, and no functional version exists today. Others are copies of a gene that has so evolved over time, the pseudogene's sequence may reflect an older, earlier version of its parent. Consequently, intergenic regions can be seen as vast molecular fossil beds offering a silent record of events in our evolutionary past.

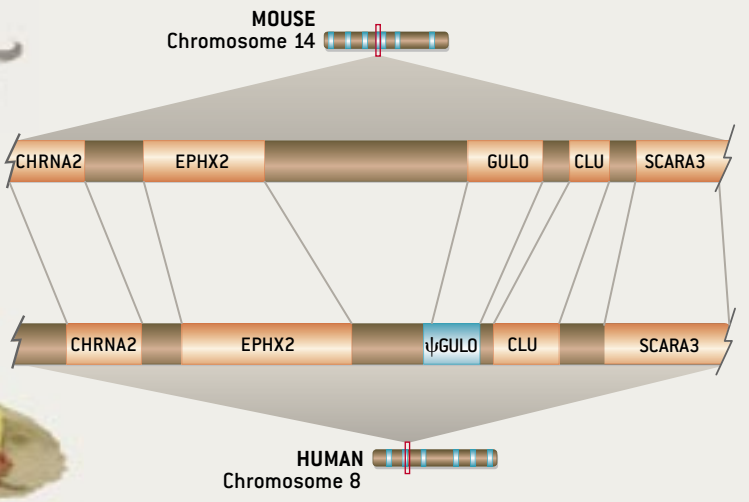
Family Histories

THE PRINCIPLES of natural selection appear to extend to individual genes, strongly constraining mutations in the sequences of functional genes. Beneficial mutations in a gene that improve the organism's fitness therefore tend to be preserved, whereas a sequence change that impairs a gene's function leads it to be discarded.

Once consigned to the genomic junk pile, however, pseudogenes are released from this selection pressure and are free to accumulate all kinds of mutations, including changes that would be deleterious to normal genes. Scientists can use this tendency to derive a kind of molecular clock from the nucleotide changes in pseudogenes and use it to study the overall dynamics and evolution of the genome. Tracking the evolutionary path of genes and pseudogenes helps molecular biologists to uncover instances of gene birth and death just as the study of mineral fossils tells paleontologists about the creation and extinction of species.

Our group has surveyed pseudogenes in the genomes of many forms of life,

CHROMOSOMES of humans and mice carry a very similar array of functional genes (*orange*) but reveal distinct differences in their pseudogenes (*blue*), which can highlight important turning points in an organism's evolutionary history. For example, the counterpart of a mouse gene called *Gulo* has become a pseudogene (ΨGulo) in humans and other primates. *Gulo* makes an enzyme that is the last element in a biochemical pathway for synthesizing vitamin C. Most mammals possess the active gene, but the primate lineage seems to have lost it more than 40 million years ago. When the *Gulo* gene became a pseudogene, primates became dependant on food sources of vitamin C to avoid scurvy.



LUCY READING-IKKANDA (*illustrations*); JANE BURTON *Dorling Kindersley/Getty Images* (*mouse*); PAUL BARTON *zefa/Corbis* (*baby*)

Differences in pseudogenes offer hints about diverse life histories.

ranging from bacteria to more complex organisms, such as yeast, worms, flies and mice, and their prevalence across a wide range of creatures is striking. The number of pseudogenes in different genomes varies greatly, more so than genes, and it is not readily predictable, because it is neither strictly proportional to the size of a genome nor to the total number of genes. Comparisons of pseudogenes in related genomes can nonetheless reveal important information about the history of specific genes and the general workings of molecular evolution.

One of the largest known gene families in mammals, for example, consists of more than 1,000 different genes encoding olfactory receptors, the cell-surface proteins that confer our sense of smell. Detailed analyses of olfactory receptor (OR) genes and pseudogenes by Doron Lancet and Yoav Gilad of the Weizmann Institute of Science in Rehovot, Israel, show that humans have lost a large number of functional olfactory receptor genes during evolution, and we now have fewer than 500 of them in our genome. For comparison, versions of about 300 human olfactory receptor pseudogenes are still functional genes in the genomes of rats and mice.

This difference is not surprising given that most animals depend more for their survival on the sense of smell than humans do. In fact, humans have considerably more olfactory receptor pseudogenes than chimpanzees do, indicating that we lost many of those functional genes after our split from the ape lineage. Apes, however, have a higher proportion of olfactory receptor pseudogenes (30 to 40 percent of the OR family) than rodents or dogs do, suggesting that some influence has allowed the entire ape lineage to get by with a somewhat reduced sense of smell.

Lancet and his colleagues found in studies of apes, monkeys and other distant primate cousins that the greatest loss of olfactory receptor genes—that is, the greatest increase in OR pseudogenes—occurred in ape and monkey lineages that evolved the ability to see color in three wavelengths of visible light. The link may suggest that a sensory trade-off took place over time in the primate lineage when better eyesight made an acute sense of smell less critical.

Often, genes involved in an organism's response to its environment are subject to extensive duplication and diversification over time, leading to large gene families, such as the olfactory receptor repertoire. Many dead-on-arrival pseudogene copies are an immediate byproduct of this process. But the subsequent death of additional duplicates, which gives rise to new pseudogenes, is also frequently connected to changes in an organism's environment or its circumstances. Consequently, differences in the pseudogenes of animals offer hints about their diverse life histories that are not as easily detected in comparisons of working genes, which are strongly constrained by their function.

Analysis of the mouse genome, for example, has shown that 99 percent of human genes have a corresponding version in the mouse. Although the human and mouse lineages diverged some 75 million years ago, nearly all of the human genome can be lined up against equivalent regions, known as syntenic blocks, in the mouse genome. Yet despite this similarity in functional genes and overall genome structure, just a small fraction of the known human pseudogenes have an obvious counterpart in the mouse.

What is more, some of the specific gene families giving rise to pseudogenes differ significantly between mouse and human. Using the rate of sequence decay relative to the parent genes to determine their age, it is also clear that many pseudogenes in the human and mouse genomes have arisen at different times. These observations indicate that very disparate events have led to independent bursts of retrotransposition that created pseudogenes in each of the lineages.

THE AUTHORS

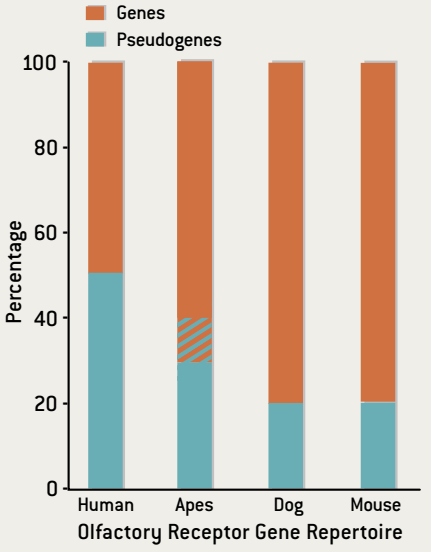
MARK GERSTEIN and *DEYOU ZHENG* are bioinformaticians at Yale University, where Gerstein is A. L. Williams Professor of Biomedical Informatics and co-director of the Yale Program in Computational Biology and Bioinformatics. Zheng, after completing his Ph.D. at Rutgers University, joined Gerstein's group in 2003 to begin investigating pseudogene activity and evolution. Both authors were initially interested in studying molecular structure and simulation, as described in Gerstein's previous article for *Scientific American* with Michael Levitt, "Simulating Water and the Molecules of Life" (November 1998). But Gerstein and Zheng were intrigued by the enormous data analysis challenges posed by the sequencing of the human genome and chose to start scanning and sifting the regions of DNA between genes.

Scanning and Sifting

STUDIES OF PSEUDOGENES in their own right are really just beginning, because these fossil genes were long viewed as little more than a nuisance. Early efforts to catalogue pseudogenes were largely driven by the need to distinguish them from true genes when annotating genome sequences. Identifying pseudogenes is not as straightforward as recognizing genes, however. Based on characteristic elements, pattern-seeking computer algorithms can scan DNA sequences and identify genes with moderate success. Recognition of pseudogenes, in contrast, relies primarily on their similarity to genes and their lack of function. Computers can detect similarity by exhaustively aligning chunks of intergenic DNA against all possible parent genes. Establishing a suspected pseudogene's inability to function is more challenging.

Just as a living organism can die of many different causes, a variety of deleterious mutations affecting any step in the process of making a protein can disable a copied gene, turning it into a pseudogene. But the sequence itself can offer clues to whether a mutation is debilitating. We can look for premature "stop" signs, as well as insertions or deletions of nucleotides that shift the reading frame of cellular machinery that decodes the gene's information for making a protein. These disablements cannot be tolerated by true genes and are thus typical manifestations of pseudogenes.

More subtly, the theory of neutral evolution introduced by mathematical biologist Motoo Kimura in the 1960s holds that nonfunctional DNA sequences can change freely, without the constraint of natural selection. Thus, individual nucleotide mutations can be divided into two types: those that would preserve the amino acid sequence of the protein encoded by a gene, known as synonymous changes, and nonsynonymous changes that would alter the meaning of the sequence. Because changing a protein's amino acid sequence can abolish its function, a gene under selective pressure will be more likely to contain synonymous mutations, whereas a nonfunctional DNA sequence will not be subject to that constraint.



CILIA projecting from human olfactory epithelium (*left*) are studded with invisible odorant molecule receptors that detect smells. A family of more than 1,000 genes encoding those olfactory receptors in mammals has been identified, although individual species vary widely in their total number of olfactory receptor genes and the percentage of that repertoire that has died and become pseudogenes. Large-scale pseudogenization is most often seen among genes that, like the olfactory receptor family, are responsible for responses to the environment. An organism's pseudogenes may therefore reflect species-specific changes in circumstances during its evolutionary history.

Comparison of pseudogenes among genomes has revealed a puzzling phenomenon, however: a few pseudogenes appear to be better preserved than one would expect if their sequences were drifting neutrally. Such pseudogenes may therefore be under evolutionary constraint, which implies that they might have some function after all. One way to try to ascertain whether pseudogenes are functioning is to see whether they are transcribed into RNA. Recent experiments by Thomas Gingeras of Affymetrix and by Michael Snyder of Yale University have found that a significant fraction of the intergenic regions in the human genome are actively transcribed. In their studies, in fact, more than half the heavily transcribed sequences map to regions outside of known genes. What is more, a number of those transcriptionally active intergenic areas overlap with pseudogenes, suggesting that some pseudogenes may have life left in them.

Our research group is part of a consortium of laboratories working to understand what is going on in the dark matter of the genome. We are now in the pilot phase of a project to create an "encyclopedia of DNA elements" (referred to as ENCODE) whose ultimate goal is to identify all of the genome's parts and their function. Previous studies as well as preliminary ENCODE data indicate that at least one tenth of the pseudogenes in the human genome are transcriptionally active. Knowing that so many pseudogenes are transcribed does not tell us their function, but together with evidence that certain pseudogenes are better preserved than background intergenic sequences, it certainly challenges the classical view of pseudogenes as dead.

One possibility is that pseudogenes play some ongoing part in regulating the activity of functional genes. Molecular biologists have come to understand in recent years that many genes in higher organisms do not code for a final protein product, but instead their RNA transcripts act to control other genes. These regulatory RNA molecules can variously activate or repress another gene or can interfere with the translation of that

RICHARD COSTANZO *Virginia Commonwealth University (micrograph)*; LUCY READING-IKKANDA *(illustrations)*

Nature may have figured out a way to reuse the broken parts of genes.

gene's mRNA transcript into a functional protein. And at least two examples of pseudogenes behaving in a similar manner have been documented so far.

The first was reported in 1999 by Michael O'Shea's research group at the University of Sussex in England. The investigators found that in the neurons of a common pond snail, both the gene for nitric oxide synthase (NOS) and its related pseudogene are transcribed into RNA but that the RNA transcript of the NOS pseudogene inhibits protein production from the transcript of the normal NOS gene.

Then, in 2003, Shinji Hirotsune of the Saitama Medical School in Japan traced deformities in a group of experimental baby mice to the alteration of a pseudogene. The inactivity of an important regulatory gene called *Makorin1* had derailed the development of the mice, but Hirotsune had not done anything to *Makorin1*. He had accidentally disrupted the Makorin1 pseudogene, which affected the function of its counterpart, the *Makorin1* gene.

Perhaps two dozen examples of specific pseudogenes that appear to be active in some way—often only in certain cells of an organism—have been identified, although the findings are still preliminary. Because many pseudogenes have sequences highly similar to those of their parent genes, it is very tempting to speculate that the NOS and Makorin1 pseudogenes are not just isolated cases. Yet it is hard to imagine that these two pseudogenes had the specific roles they now perform when they first arose. Instead their activity may be the result of selection preserving happy accidents or of nature having figured out an efficient way to reuse the broken parts of genes by converting them into regulatory elements.

Protogenes

AN EXCITING ERA of molecular paleontology is just beginning. We have barely scratched the surface of the pseudogene strata, and once we drill deeper, the number of identified pseudogenes will most likely grow and we may find more surprises. Large-scale pseudogene identification is a very dynamic data-mining process. Current techniques rely heavily on sequence comparison to well-characterized genes, and although they can readily identify recently generated pseudogenes, very ancient and decayed sequences are probably escaping detection. As the sequence and annotation of the human genome itself are refined and updated, characterization of pseudogenes will improve as well.

Recent hints that not all pseudogenes are entirely dead have been intriguing, and some evidence also exists, for the possibility of pseudogene resurrection—a dead gene turning back into a living one that makes a functional protein product. Careful sequence comparisons have shown that one cow gene for a ribonuclease enzyme was a pseudogene for much of its history but appears to have been reactivated during recent evolutionary time. Slight differences in the pseudogene complements of individual people have also been found—for example, a few olfactory receptor pseudogenes straddle the fence: in most people they are pseudogenes, but in some they are intact, working genes. These anomalies could arise if random mutation reverses the disablement that originally produced the pseudogene. Might they account for individuals' differing sensitivities to smell? Perhaps, although it is too early to guess at the scope or significance of this unexpected source of genetic variation among humans.

Our studies have suggested, however, that in yeast, certain cell-surface protein pseudogenes are reactivated when the organism is challenged by a stressful new environment. Thus, pseudogenes may be considered not only as dead genes (which nonetheless provide fascinating new insights into our past) but also as potentially unborn genes: a resource tucked away in our genetic closet to be drawn on in changing circumstances, one whose possible roles in our present and future genomes are just beginning to unfold. SA

MORE TO EXPLORE

Pseudogenes: Are They "Junk" or Functional DNA? Evgeniy S. Balakirev and Francisco J. Ayala in *Annual Review of Genetics,* Vol. 37, pages 123–151; December 2003.

Human Specific Loss of Olfactory Receptor Genes. Yoav Gilad, Doron Lancet et al. in *Proceedings of the National Academy of Sciences,* Vol. 100, No. 6, pages 3324–3327; March 18, 2003.

Large-Scale Analysis of Pseudogenes in the Human Genome. Zaolei Zhang and Mark Gerstein in *Current Opinion in Genetics & Development,* Vol. 14, No. 4, pages 328–335; August 2004.

www.pseudogene.org/