The New York Times
nytimes.com

November 11, 2008

Now: The Rest of the Genome

By CARL ZIMMER

Over the summer, Sonja Prohaska decided to try an experiment. She would spend a day without ever saying the word “gene.” Dr. Prohaska is a bioinformatician at the University of Leipzig in Germany. In other words, she spends most of her time gathering, organizing and analyzing information about genes. “It was like having someone tie your hand behind your back,” she said.

But Dr. Prohaska decided this awkward experiment was worth the trouble, because new large-scale studies of DNA are causing her and many of her colleagues to rethink the very nature of genes. They no longer conceive of a typical gene as a single chunk of DNA encoding a single protein. “It cannot work that way,” Dr. Prohaska said. There are simply too many exceptions to the conventional rules for genes.

It turns out, for example, that several different proteins may be produced from a single stretch of DNA. Most of the molecules produced from DNA may not even be proteins, but another chemical known as RNA. The familiar double helix of DNA no longer has a monopoly on heredity. Other molecules clinging to DNA can produce striking differences between two organisms with the same genes. And those molecules can be inherited along with DNA.

The gene, in other words, is in an identity crisis.

This crisis comes on the eve of the gene’s 100th birthday. The word was coined by the Danish geneticist Wilhelm Johanssen in 1909, to describe whatever it was that parents passed down to their offspring so that they developed the same traits. Johanssen, like other biologists of his generation, had no idea what that invisible factor was. But he thought it would be useful to have a way to describe it.

“The word ‘gene’ is completely free from any hypothesis,” Johanssen declared, calling it “a very applicable little word.”

Over the next six decades, scientists transformed that little word from an abstraction to concrete reality. They ran experiments on bread mold and bacteria, on fruit flies and corn. They discovered how to alter flowers and eyes and other traits by tinkering with molecules inside cells. They figured out that DNA was a pair of strands twisted around each other. And by the 1960s, they had a compelling definition of the gene.

A gene, they argued, was a specific stretch of DNA containing the instructions to make a protein molecule. To make a protein from a gene, a cell had to read it and build a single-stranded copy known as a transcript out of RNA. This RNA was then grabbed by a cluster of molecules called a ribosome, which used it as a template to build a protein.

A gene was also the fundamental unit of heredity. Every time a cell divided, it replicated its genes, and parents passed down some of their genes to their offspring. If you inherited red hair — or a predisposition for breast cancer — from your mother, chances were that you inherited a gene that helped produce that trait.

This definition of the gene worked spectacularly well — so well, in fact, that in 1968 the molecular biologist Gunther Stent declared that future generations of scientists would have to content themselves with "a few details to iron out."

The Details

Stent and his contemporaries knew very well that some of those details were pretty important. They knew that genes could be shut off and switched on when proteins clamped onto nearby bits of DNA. They also knew that a few genes encoded RNA molecules that never became proteins. Instead, they had other jobs, like helping build proteins in the ribosome.

But these exceptions did not seem important enough to cause scientists to question their definitions. "The way biology works is different from mathematics," said Mark Gerstein, a bioinformatician at Yale. "If you find one counterexample in mathematics, you go back and rethink the definitions. Biology is not like that. One or two counterexamples — people are willing to deal with that."

More complications emerged in the 1980s and 1990s, though. Scientists discovered that when a cell produces an RNA transcript, it cuts out huge chunks and saves only a few small remnants. (The parts of DNA that the cell copies are called exons; the parts cast aside are introns.) Vast stretches of noncoding DNA also lie between these protein-coding regions. The 21,000 protein-coding genes in the human genome make up just 1.2 percent of that genome.

The Genome

In 2000, an international team of scientists finished the first rough draft of that genome — all of the genetic material in a human cell. They identified the location of many of the protein-coding genes, but they left the other 98.8 percent of the human genome largely unexplored.

Since then, scientists have begun to wade into that genomic jungle, mapping it in fine detail.

One of the biggest of these projects is an effort called the Encyclopedia of DNA Elements, or Encode for short. Hundreds of scientists are carrying out a coordinated set of experiments to determine the function of every piece of DNA in the human genome. Last summer they published their results on 1 percent of the genome — some three million "letters" of DNA. The genetic code is written in letters, like the title of the movie "Gattaca," with each letter standing for a molecule called a base: guanine (G), adenine (A), thymine (T), cytosine (C). The Encode team expects to have initial results on the other 99 percent by next year.

Encode's results reveal the genome to be full of genes that are deeply weird, at least by the traditional standard of what a gene is supposed to be. "These are not oddities — these are the rule," said Thomas R. Gingeras of Cold Spring Harbor Laboratory and one of the leaders of Encode.

A single so-called gene, for example, can make more than one protein. In a process known as alternative splicing, a cell can select different combinations of exons to make different transcripts. Scientists identified the first cases of alternative splicing almost 30 years ago, but they were not sure how common it was. Several studies now show that almost all genes are being spliced. The Encode team estimates that the average protein-coding region produces 5.7 different transcripts. Different kinds of cells appear to produce different transcripts from the same gene.

Even weirder, cells often toss exons into transcripts from other genes. Those exons may come from distant locations, even from different chromosomes.

So, Dr. Gingeras argues, we can no longer think of genes as being single stretches of DNA at one physical location.

“I think it’s a paradigm shift in how we think the genome is organized,” Dr. Gingeras said.

The Epigenome

But it turns out that the genome is also organized in another way, one that brings into question how important genes are in heredity. Our DNA is studded with millions of proteins and other molecules, which determine which genes can produce transcripts and which cannot. New cells inherit those molecules along with DNA. In other words, heredity can flow through a second channel.

One of the most striking examples of this second channel is a common flower called toadflax. Most toadflax plants grow white petals arranged in a mirror-like symmetry. But some have yellow five-pointed stars. These two forms of toadflax pass down their flower to their offspring. Yet the difference between their flowers does not come down to a difference in their DNA.

Instead, the difference comes down to the pattern of caps that are attached to their DNA. These caps, made of carbon and hydrogen, are known as methyl groups. The star-shaped toadflax have a distinct pattern of caps on one gene involved in the development of flowers.

DNA is not just capped with methyl groups; it is also wrapped around spool-like proteins called histones that can wind up a stretch of DNA so that the cell cannot make transcripts from it. All of the molecules that hang onto DNA, collectively known as epigenetic marks, are essential for cells to take their final form in the body. As an embryo matures, epigenetic marks in different cells are altered, and as a result they develop into different tissues. Once the final pattern of epigenetic marks is laid down, it clings stubbornly to cells. When cells divide, their descendants carry the same set of marks. “They help cells remember what genes to keep on, and what genes can never be turned on,” said Bradley Bernstein of Harvard University.

Scientists know much less about this “epigenome” than the genome. In September, the National Institutes of Health began a $190 million program to start mapping epigenetic marks on DNA in different tissues. “Now we can chart all these changes beyond the gene,” said Eric Richards of Cornell University.

This survey may provide clues to the origins of cancer and other diseases. It has long been known that when DNA mutates, a cell may become prone to turning cancerous. Some studies now suggest that when epigenetic marks are disturbed, cells may also be made more vulnerable to cancer, because essential genes are shut off and genes that should be shut off are turned on. What makes both kinds of changes particularly dangerous is that they are passed down from a cell to all its descendants.

When an embryo begins to develop, the epigenetic marks that have accumulated on both parents’ DNA are stripped away. The cells add a fresh set of epigenetic marks in the same pattern that its parents had when they were embryos.

This process turns out to be very delicate. If an embryo experiences certain kinds of stress, it may fail to lay

down the right epigenetic marks.

In 1944, for example, the Netherlands suffered a brutal famine. Scientists at the University of Leiden recently studied 60 people who were conceived during that time. In October, the researchers reported that today they still have fewer epigenetic marks than their siblings. They suggest that during the 1944 famine, pregnant mothers could not supply their children with the raw ingredients for epigenetic marks.

In at least some cases, these new epigenetic patterns may be passed down to future generations. Scientists are debating just how often this happens. In a paper to be published next year in The Quarterly Review of Biology, Eva Jablonski and Gal Raz of Tel Aviv University in Israel assemble a list of 101 cases in which a trait linked to an epigenetic change was passed down through three generations

For example, Matthew Amway of Washington State University and his colleagues found that exposing pregnant rats to a chemical for killing fungus disrupted the epigenetic marks in the sperm of male embryos. The embryos developed into adult rats that suffered from defective sperm and other disorders, like cancer. The males passed down their altered epigenetic marks to their own offspring, which passed them down to yet another generation.

Last year Dr. Amway and his colleagues documented an even more surprising effect of the chemical. Female rats exposed in the womb avoided mating with exposed male rats. The scientists found this preference lasted at least three generations.

While these experiments are eye-opening, scientists are divided about how important these generation-spanning changes are. “There’s a lot of disagreement about whether it matters,” Dr. Richards said.

RNA in the Spotlight

Epigenetic marks are intriguing not just for their effects, but also for how they are created in the first place. To place a cap of methyl groups on DNA, for example, a cluster of proteins must be guided to the right spot. It turns out they must be led there by an RNA molecule that can find it.

These RNA guides, like the RNA molecules in ribosomes, do not fit the classical concept of the gene. Instead of giving rise to a protein, these RNA molecules immediately start to carry out their own task in the cell. Over the last decade, scientists have uncovered a number of new kinds of RNA molecules that never become proteins. (Scientists call them noncoding RNA.) In 2006, for example, Craig Mello of the University of Massachusetts and Andrew Fire of Stanford University won the Nobel Prize for establishing that small RNA molecules could silence genes by interfering with their transcription.

These discoveries left scientists wondering just how much noncoding RNA our cells make. The early results of Encode suggest the answer is a lot. Although only 1.2 percent of the human genome encodes proteins, the Encode scientists estimate that a staggering 93 percent of the genome produces RNA transcripts.

John Mattick, an Encode team member at the University of Queensland in Australia, is confident that a lot of those transcripts do important things that scientists have yet to understand. “My bet is the vast majority of it — I don’t know whether that’s 80 or 90 percent,” he said.

“When you cross the Rubicon and look back, you see the protein-centric view as being quite primitive,” he

said.

Certain versions of those RNA-coding genes may raise the risk of certain diseases. As part of the Encode project, scientists identified the location of variations in DNA that have been linked to common diseases like cancer. A third of those variations were far from any protein-coding gene. Understanding how noncoding RNA works may help scientists figure out how to use drugs to counteract genetic risks for diseases. “This is going to be a huge topic of research this coming decade,” said Ewan Birney, one of the leaders of the Encode project at the European Bioinformatics Institute.

Despite the importance of noncoding RNA, Dr. Birney suspects that most of the transcripts discovered by the Encode project do not actually do much of anything. “I think it’s a hypothesis that has to be on the table,” he said.

David Haussler, another Encode team member at the University of California, Santa Cruz, agrees with Dr. Birney. “The cell will make RNA and simply throw it away,” he said.

Dr. Haussler bases his argument on evolution. If a segment of DNA encodes some essential molecule, mutations will tend to produce catastrophic damage. Natural selection will weed out most mutants. If a segment of DNA does not do much, however, it can mutate without causing any harm. Over millions of years, an essential piece of DNA will gather few mutations compared with less important ones.

Only about 4 percent of the noncoding DNA in the human genome shows signs of having experienced strong natural selection. Some of those segments may encode RNA molecules that have an important job in the cell. Some of them may contain stretches of DNA that control neighboring genes. Dr. Haussler suspects that most of the rest serve no function.

“Most of it is baggage being dragged along,” he said.

But the line between the useless baggage and the useful DNA is hard to draw. Mutations can make it impossible for a cell to make a protein from a gene. Scientists refer to such a disabled piece of DNA as a pseudogene. Dr. Gerstein and his colleagues estimate that there are 10,000 to 20,000 pseudogenes in the human genome. Most of them are effectively dead, but a few of them may still make RNA molecules that serve an important function. Dr. Gerstein nicknames these functioning pseudogenes “the undead.”

Alien DNA

Much of the baggage in the genome comes not from dead genes, however, but from invading viruses. Viruses repeatedly infected our distant ancestors, adding their DNA to the genetic material passed down from generation to generation. Once these viruses invaded our genomes, they sometimes made new copies of themselves, and the copies were pasted in other spots in the genome. Over many generations, they mutated and lost their ability to move.

“Our genome is littered with the rotting carcasses of these little viruses that have made their home in our genome for millions of years,” Dr. Haussler said.

As these chunks of viral DNA hop around, they can cause a lot of harm. They can disrupt the genome, causing it to stop making essential proteins. Hundreds of genetic disorders have been linked to their leaps.

One of the most important jobs that noncoding RNA serves in the genome is preventing this virus DNA from spreading quickly.

Yet some of these invaders have evolved into useful forms. Some stretches of virus DNA have evolved to make RNA genes that our cells use. Other stretches have evolved into sites where our proteins can attach and switch on nearby genes. “They provide the raw material for innovation,” Dr. Haussler said.

In this jungle of invading viruses, undead pseudogenes, shuffled exons and epigenetic marks, can the classical concept of the gene survive? It is an open question, one that Dr. Prohaska hopes to address at a meeting she is organizing at the Santa Fe Institute in New Mexico next March.

In the current issue of American Scientist, Dr. Gerstein and his former graduate student Michael Seringhaus argue that in order to define a gene, scientists must start with the RNA transcript and trace it back to the DNA. Whatever exons are used to make that transcript would constitute a gene. Dr. Prohaska argues that a gene should be the smallest unit underlying inherited traits. It may include not just a collection of exons, but the epigenetic marks on them that are inherited as well.

These new concepts are moving the gene away from a physical snippet of DNA and back to a more abstract definition. “It’s almost a recapture of what the term was originally meant to convey,” Dr. Gingeras said.

A hundred years after it was born, the gene is coming home.