Name _____

Human Genome Articles 1-3

1. According to the public press, the recently completed Human Genome Project (HGP) boiled down to a bitter and hard-fought RACE. Who was "racing" who ... compare and contrast their sequencing techniques, and in your opinion ... who won the race ... Explain?

2. Explain the difference between the human genome and the human proteome. Which type of data (genome vs. proteome) do scientists predict will have the most significant impact on 21st century medicine ... Explain?

3. Explain how the human genome map released in 2007 differs from the original map completed in 2003. Summarize new insights that this has allowed us to understand about our genome.

GENOME

Showdown In the DNA Corral #1A

Rival teams of gene sleuths are still bitter

BY SHARON BEGLEY

THE COMPLETION OF THE HUMAN genome sequence, announced last week, promises to usher in a new age of biology. In these FAQs, we discuss how the genome promises drugs targeted to individual DNA; early warnings of diseases our genes put us at risk for; a deeper understanding of human evolution ...

Sure, sure. But who won?
You mean the public consortium, financed by the U.S. government and Britain's Wellcome Trust, or Celera, which announced in 1998 that it would reach the finish line first? But both sides denied it was a race.

And, who won?
Well, the public project had sequenced less than 5 percent of the genome when Celera leaped in. Celera's J. Craig Venter said he'd finish in less than three years, which lit a fire under his rivals. But, most important, Venter also said he would do it through a faster, cheaper method called whole-genome shotgun sequencing. His claim last week to have done so "is completely false," says John McPherson of Washington University, a leader of the public project. It looks like Celera could not assemble the full genome sequence through shotgunning alone. Instead, it had to use the public consortium's "maps," which Celera had derided as wasteful, expensive and slow. Says one scientist in the public project, "The shotgun did not work." And since Celera used public-project data, to critics it seems as if the company crossed the finish line only because it attached a tow rope to the race car in front of it.

Does Celera agree?
What a sense of humor you have. Venter says he "shouldn't even dignify [the criticism] with a response." But he does. The ...un method "works spectacularly," he ...increasing sequencing speeds by an order of magnitude." He concedes, though, that "we used all the information we could find"—meaning data from the public project—"to validate our construction and align the pieces [of genome sequence] on the right chromosomes."

A fight to the finish—and beyond: *Venter (top) and Collins*

Why do passions run so high?
Because, innocent friend, the public project's scientists, led by Francis Collins, feel they "had to put up with three years of this crap, being told that Celera was going to do it faster and better," as one said. Celera, on the other hand, feels "our critics will never be happy as long as anything we're doing is successful," Venter says.

I hope I won't regret this, but what is the difference between the two approaches?
It's not that complicated. Celera shreds the entire genome—all 3.2 billion chemical letters. Think of it as shredding a 23-volume encyclopedia. Sequencing machines determine the order of the chemicals (denoted A, T, C or G) in each fragment, but the machine can't sequence anything much longer than 500 letters. So you sequence one fragment of 500 or so, then another and another. The result is what Eric Lander of the Whitehead Institute calls "tossed genome salad: you don't know what order [each batch of 500 letters] goes in." It's like knowing the letters in the words in isolated sentences of your encyclopedia, but having no idea what order the pages go in. Venter thought his computers and algorithms would assemble the fragments correctly. But Celera had to resort to the public project's maps for that.

And a map is ...
It's sort of a way to keep track of which page of the encyclopedia your letters came from. Instead of having one pile of millions of shredded fragments, you have many piles, with fewer fragments. That makes assembling it all in the correct order easier. So, to oversimplify, if you see a fragment that ends with "manylettersATTGCTTTGG," and another that begins with "ATTGCTTTGGmoreletters," they probably overlap. The assembled stretch is "manyletters-ATTGCTTTGGmoreletters."

So why doesn't that work with the zillions of fragments that the shotgun gives you?
Because the human genome is about 50 percent repetitive. If you have hundreds of those ATTGCTTTGGs, it's tough to figure what overlaps with what. "So many parts of the genome look exactly alike, there's a real risk of sticking wrong parts together," says McPherson. Celera may have the last laugh, though: in just three days, more than 1 million users accessed its genome assembly. Although some looks are free, other uses cost upwards of $15,000 a year. ■

34,000
... or so genes in the human genome, which has about 3.2 billion chemical 'letters.'

19,099
... in the roundworm genome, completed in 1998. It has 97 million A's, T's, C's and G's.

13,601
... in the fruit-fly genome, completed in 2000. It has 185 million of those chemical letters.

TOP TO BOTTOM: SHAHAR AZRAN—CORBIS SYGMA, CHUCK KENNEDY—KRT, KENNETH EWARD—PHOTO RESEARCHERS

#1B

12.18 The Human Genome Project revealed that most of the human genome does not consist of genes

The most ambitious genomics project to date has been the **Human Genome Project (HGP).** The goals of the HGP were to determine the nucleotide sequence of all DNA in the human genome and to identify the location and sequence of every gene. The HGP began in 1990 as an effort by a consortium of 20 government-funded research centers in six countries. Several years into the project, private companies, chiefly Celera Genomics in the United States, joined the effort. At the completion of the final draft of the sequence in 2004, over 99% of the genome had been determined to 99.999% accuracy. As of 2007, there remain a few hundred gaps of unknown sequences within the human genome that will require special methods to figure out. The DNA sequences determined by the HGP have been deposited in a database, called Genbank, that is publically available via the Internet.

The chromosomes in the human genome (22 autosomes plus the X and Y sex chromosomes) contain approximately 3.2 billion nucleotide pairs of DNA. To try to get a sense of this quantity of DNA, imagine that its nucleotide sequence is printed in letters (A, T, C, and G) like the letters in this book. At this size, the sequence would fill a stack of books 18 stories high! The biggest surprise from the HGP is the small number of human genes. The current estimate is about 21,000 genes—very close to the number found in the nematode worm. How, then, do we account for human complexity? Part of the answer may lie in alternative RNA splicing (see Module 11.6); scientists think that a typical human gene probably specifies several polypeptides.

In humans, like most complex eukaryotes, only a small amount of our total DNA (about 1.5%) is contained in genes that code for proteins, tRNAs, or rRNAs (Figure 12.18). Most multicellular eukaryotes have a huge amount of noncoding DNA—about 98.5% of human DNA is of this type. Some noncoding DNA is made up of gene control sequences such as promoters and enhancers (see Chapter 11). The remaining noncoding DNA has been dubbed "junk DNA," a tongue-in-cheek way of saying that scientists don't fully understand its functions.

Much of the DNA between genes consists of repetitive DNA, nucleotide sequences present in many copies in the genome. Some of this noncoding DNA is used to create DNA profiles by STR analysis, as discussed in Module 12.14. Stretches of DNA with thousands of such repetitions are also prominent at the centromeres and ends of chromosomes—called **telomeres**—suggesting that this DNA plays a role in chromosome structure.

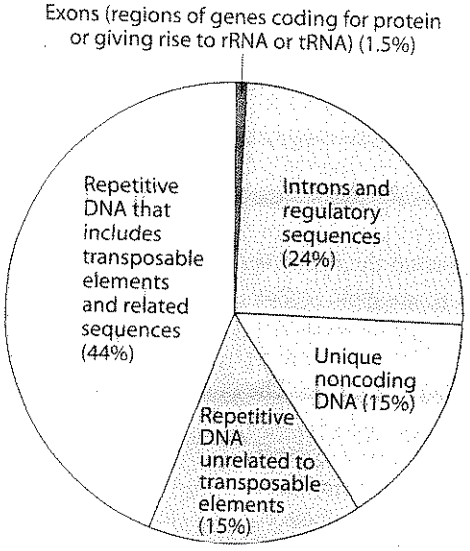


Figure 12.18 Composition of the human genome

In the second main type of repetitive DNA, each repeated unit is hundreds of nucleotides long, and the copies are scattered around the genome. Most of these sequences seem to be associated with **transposable elements** ("jumping genes"), DNA segments that can move or be copied from one location to another in a chromosome and even between chromosomes. Researchers believe that transposable elements, through their copy-and-paste mechanism, are responsible for the proliferation of dispersed repetitive DNA in the human genome.

The potential benefits of having a complete map of the human genome are enormous. For instance, hundreds of disease-associated genes have already been identified. One example is the gene that is mutated in an inherited type of Parkinson's disease, a debilitating brain disorder that causes tremors of increasing severity. Until recently, Parkinson's disease was thought to have only an environmental basis; there was no evidence of a hereditary component. But data from the Human Genome Project mapped a very small number of cases of Parkinson's disease to a specific gene. Interestingly, an altered version of the protein encoded by this gene has also been tied to Alzheimer's disease, suggesting a link between these two brain disorders. Moreover, the same gene is also found in rats, where it plays a role in the sense of smell, and in zebra finches, where it is thought to be involved in song learning. Cross-species comparisons such as these may uncover clues about the role played by the normal version of the protein in the brain. And such knowledge could eventually lead to treatment for the half a million Americans with Parkinson's disease.

One interesting question about the Human Genome Project is: *Whose genome* was sequenced? The answer is no one's—or at least not any one person's. The human genome sequenced by the public consortium was actually a reference genome compiled from a group of individuals. The genome sequenced by Celera consisted primarily of DNA sampled from the company's president. These representative sequences will serve as standards so that comparisons of individual differences and similarities can be made. Eventually, as the amount of sequence data multiplies, the small differences that account for individual variation within our species will come to light.

? The human genome consists of about _____ nucleotides and _____ genes spread over _____ different chromosomes (provide three numbers).

3.2 billion . . . 21,000 . . . 24 (22 autosomes plus two sex chromosomes)

1C

12.19 The whole-genome shotgun method of sequencing a genome can provide a wealth of data quickly

Sequencing an entire genome is a complex task that requires careful work. The Human Genome Project proceeded through three stages that provided progressively more detailed views of the human genome. First, geneticists combined pedigree analyses of large families to map over 5,000 genetic markers (mostly RFLPs) spaced throughout all of the chromosomes. The resulting low-resolution *linkage map* provided a framework for mapping other markers and for arranging later, more detailed maps of particular regions. Next, researchers determined the number of base pairs between the markers in the linkage map. These data helped them construct a *physical map* of the human genome. Finally came the most arduous part of the project: determining the nucleotide sequences of the set of DNA fragments that had been mapped. Advances in automated DNA sequencing were crucial to this endeavor.

This three-stage approach is logical and thorough. However, in 1992, molecular biologist J. Craig Venter proposed an alternative strategy called the **whole-genome shotgun method** and set up the company Celera Genomics to implement it. His idea was essentially to skip the genetic and physical mapping stages and start directly with the sequencing step. In the whole genome shotgun method, an entire genome is chopped by restriction enzymes into fragments that are cloned and sequenced in just one stage (Figure 12.19). High-performance computers running specialized mapping software can assemble the millions of overlapping short sequences into a single continuous sequence.

Today, the whole-genome shotgun approach is the method of choice for genomic researchers because it is fast and relatively inexpensive. However, recent research has revealed some limitations of this method, suggesting that a hybrid approach that combines whole genome shotgunning with physical or genetic maps may prove to be the most useful method in the long run.

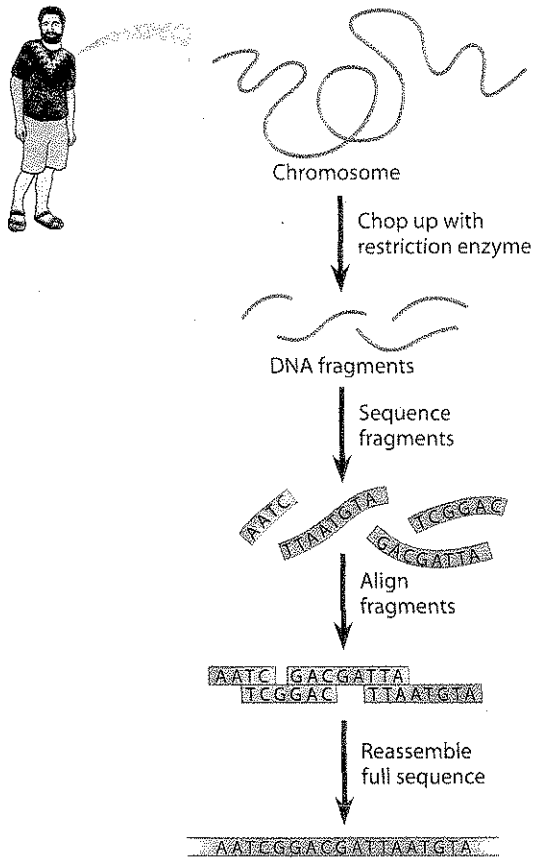


Figure 12.19 The whole-genome shotgun method

? What is the primary advantage of the whole-genome shotgun method?

■ It is faster and cheaper than the three-stage method of genome sequencing.

12.20 Proteomics is the scientific study of the full set of proteins encoded by a genome

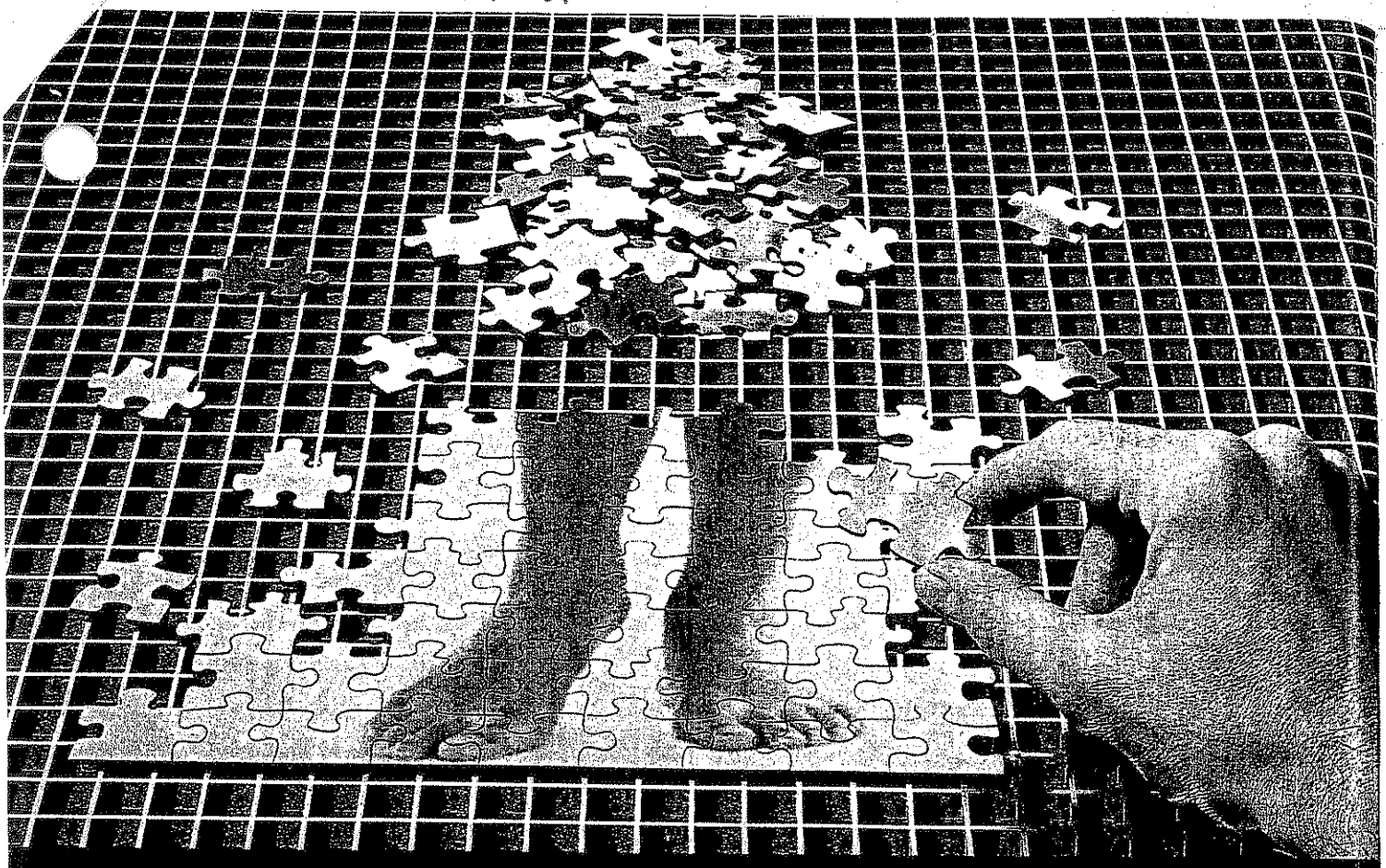
The successes in the field of genomics have encouraged scientists to attempt similar systematic study of the full protein sets (proteomes) encoded by genomes, an approach called **proteomics.** The number of proteins in humans far exceeds the number of genes (about 100,000 proteins versus about 21,000 genes). And since proteins, not genes, actually carry out most of the activities of the cell, scientists must study when and where proteins are produced in an organism and how they interact in order to understand the functioning of cells and organisms. Assembling and analyzing proteomes pose many experimental challenges, but ongoing advances are providing the tools to meet those challenges.

Genomics and proteomics are enabling biologists to approach the study of life from an increasingly holistic perspective. Biologists are now in a position to compile catalogs of genes and proteins—that is, a listing of all the "parts" that contribute to the operation of cells, tissues, and organisms. With such catalogs in hand, researchers are shifting their attention from the individual parts to how they function together in biological systems.

? If every protein is encoded by a gene, how can humans have many more proteins than genes? *Hint:* See Module 11.6.

■ The mRNA from one gene may be spliced several different ways to produce several different proteins.

2



GENETICS

Solving the Next Genome Puzzle

Identifying all our DNA was the easy part. Now, bring on the proteome.

BY SHARON BEGLEY

EVEN IN THE HIGH-MINDED FIELD of genetics, the scientists who labor to decipher genomes are, after all, only human. So, thrilled as they are to read the complete sequence of DNA in an organism—its "book of life"—a cruder puzzle also piques their curiosity. As teams of researchers sequenced more than a dozen genomes in the last few years, knocking off the yeast and rice and fruit fly and, last summer, the human, the e-mails flew fast and furious: *how big is yours?* How many genes does your genome have? For the human one, the answer is, well, not very many. After years of guesstimating that humans have something on the order of 100,000 genes, it turns out we have a humbling 34,000 or so. A roundworm has 19,099; a fruit fly has 13,601; even the mustard plant, for Pete's sake, has about 25,000. "The great abiding mystery of the human genome is how we manage to be so complex with so few genes," says Eric Lander of the Whitehead Institute, a leader of the genome project.

The two long-awaited analyses of the human genome—one by the publicly funded consortium and one by its bitter rival, biotech firm Celera—are finally being released this week, with bells-and-whistles press conferences in five cities on three continents. But industry sources and other scientists got an advance look. The human genome, they say, holds a wealth of information and several surprises. One of the oddest is that, over the eons, hundreds of genes insinuated their way into the human genome from bacteria, probably after a bacterium infected one of our distant vertebrate ancestors and slipped its DNA into its host's. These alien genes are now part of us, some performing important functions (one is an enzyme that processes brain chemicals) and some not. Analysis of the human genome also shows that the mutation rate in sperm is more than twice what it is in eggs, as David Page of the Whitehead estimated even before the genome was completely sequenced. Since mutation is the raw material of evolution, it seems that one half of humankind has been doing the heavy lifting of pulling us up from the primordial ooze. And out of the 3 billion chemical letters in the genome—those now famous A's, T's, C's and G's—there are so few variations that people the world over, from a sumo wrestler to Britney Spears, are 99.95 percent identical. But perhaps the biggest surprise is that little matter of size, and *Homo sapiens'* lack thereof. Which leads us to why, even with this wealth of new scientific knowledge, many biologists see the sequencing of the human genome as the final triumph of yesterday's biology. You can almost hear them say, sighing, "Genomes? *So* 20th century."

The new game in town is the proteome. Just as "genome" means all the DNA in an organism, so "proteome" means all the proteins. "Proteomics" is the study of that collection, and if you thought the genome was complicated (it did, after all, take on the order of $1 billion to sequence it), wait till you meet the proteome. "Compared to the human genome, proteomics involves 1,000 times more data," says Caroline Kovac of IBM Life Sciences. For although the DNA in

PLAMEN PETKOV

a liver cell is identical to the DNA in a skin cell or a brain neuron or any other cell, the proteins are not. To make things really interesting, the kinds and amounts of a cell's proteins—molecules like hemoglobin and insulin, the brain chemicals dopamine and serotonin, hormones like testosterone and estrogens, as well as the countless enzymes that keep cells running—vary not just by which type of cell you're looking at. Which proteins a cell contains also depends on things such as whether it is healthy or diseased, how old it is, its stress levels and maybe even the time of day. All told, there are probably 500,000 to 1 million human proteins. Despite the challenge, researchers are tackling the proteome for good reason. When it comes to diagnosis, prognosis and treatment of disease, the dirty little secret of genomics is that "the genome is just the beginning," says Brian Chait of Rockefeller University. "What you really want to know is, in a person's 100 billion cells, what proteins are made in each?"

And for that, the genome is not enough. The genome is the set of instructions for making proteins. But knowing the instructions doesn't get you far. That's because the 34,000 or so genes in every human cell are little more than order forms. Some orders never make it to the cellular factories that produce our proteins. Some orders make proteins that fall apart soon after leaving the factory, like an automotive lemon. Some orders are so popular that the factories make millions of them. You can't tell any of that from the order forms—the genome—alone. Three genes might dispatch order forms for proteins A, B and C, but the factory seems to regard the orders as little more than polite suggestions. It will make proteins A, B and C, sure. But it will also get fancy, making AB, AC, BC, AAB, ABC and so on. This ability to mix and match, and even to accessorize (cells will also dangle little molecules of sugar or phosphate on proteins, changing their function), sets the human genome apart from all others. "From a single gene," says John Richards of the California Institute of Technology, "you can get 10 or more different proteins. A genome analysis alone won't tell you which ones."

The reason you want to identify the proteins in the first place is that rogue genes don't cause disease. Rogue proteins do. "If you really want to understand what's going on in a disease, you have to look at the proteins," says William Rich, CEO of the biotech firm Ciphergen. Alzheimer's disease shows the value of the proteome over the genome. Yes, there are half a dozen genes that increase risk of this disease. But the only unambiguous diagnosis comes from the presence, in the brain, of sticky bits of proteins called beta amyloid fragments. Ciphergen hopes its ProteinChip will detect these killer amyloids. But because there is no beta amyloid gene, you can't screen for Alzheimer's with a DNA chip. "You can never deduce what's happening in this disease with genes alone," says Rich.

> "Proteomics is standing on the shoulders of the human genome project and asking the next tough questions."
>
> —TREVOR HAWKINS, *Joint Genome Institute*

Pharmaceutical maker Merck & Co. is using Ciphergen's chip to test candidate Alzheimer's drugs. If the chip shows that drug eliminates beta amyloid tangles, t drug is a potential winner. Another biotech, Molecular Staging, sells chips that track the progression of diseases like cancer and arthritis, in which changing levels of proteins might serve as early-warning signs of a worsening condition. Millennium Predictive Medicine has identified three dozen proteins that may be markers for hard-to-diagnose ovarian cancer. A federal proteomics project is comparing proteins in normal lung, ovarian, breast and colon tissue with those in cancerous tissue. Proteins that are more abundant in cancer could be diagnosis targets, the way PSA is for prostate cancer. And if it turns out that some of those proteins let cells divide unchecked, then an antibody that bottles up and inactivates the protein might prove an effective cancer drug. Scientists at Large Scale Proteomics Corp. (LSP) and Johns Hopkins University have compiled a list of proteins that seem to mark depression, bipolar disorder and schizophrenia.

Last month LSP unveiled the first database of human proteins, a total of 15,6[illegible] proteins from 157 tissues. That's clear[illegible] just a minuscule down payment, says Leigh Anderson, president of LSP. But researchers aren't daunted. "Proteomics is standing on the shoulders of the human genome project and asking the next questions," says Trevor Hawkins, director of the Department of Energy's Joint Genome Institute. The payoff will be the stuff of the 21st century.

With ERIKA CHECK *and* ADAM ROGERS

Mixing and Matching in the Protein Factory

Humans seem to have a mere 34,000 or so genes, the instructions for making the 1 million proteins like insulin, hemoglobin and testosterone that keep us alive. How can 34,000 genes make 1 million proteins? Very cleverly.

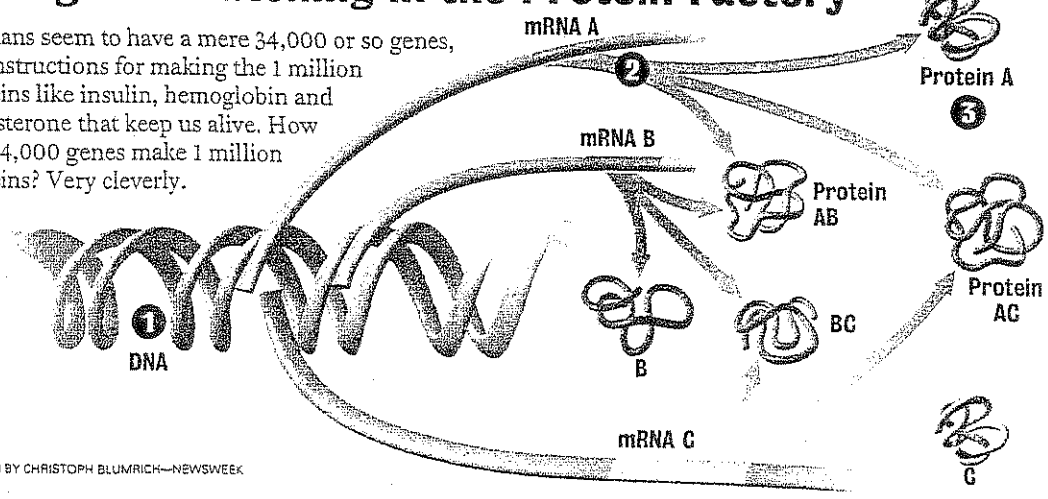


DIAGRAM BY CHRISTOPH BLUMRICH—NEWSWEEK

1 Genes in the DNA strand carry the code for making proteins

2 Each gene is copied into an mRNA molecule—here, A, B and C— that travels to the cell's protein factories

3 Depending how the mRNA is spliced, it can make protein A, protein B, protein C or some combination like AB or even ABC

3

Genome Article #3: Advance in DNA mapping offers fuller picture of genetic profile

Kansas City Star, September 9, 2007 Author: RICK WEISS, The Washington Post

WASHINGTON Scientists have for the first time determined the order of virtually every letter of DNA code in an individual. It offers an unprecedented readout of the separate genetic contributions made by that person's mother and father. By providing a detailed look at maternal and paternal DNA strands, rather than the blended composite that was yielded by the 2001 Human Genome Project, the work offers the clearest snapshot yet of just how different those two contributions can be.

Assuming that the newly decoded sequence is typical, as scientists presume it is, there are five times as many differences between individuals' DNA as was previously thought. Of more practical import, the ability to create such a detailed genetic profile with relative ease suggests that it may not be long before people of ordinary means will be able to have their complete DNA codes spelled out, scientists said. That could tell a lot about a person's health risks, because such a profile would include not only the few genes that significantly increase the likelihood of getting certain diseases but also the many "lesser" genes that pose modest risks individually but that together have the lion's share of impact on health. For better or worse, the advance also stands to bring science to the pastime of guessing which parent deserves blame or credit for passing along certain traits. "This is the ultimate form of genealogy," said Stephen Scherer of the Hospital for Sick Children in Toronto, who was part of the multimillion-dollar project described last week in the journal PLoS Biology. "You'll have incredible information about yourself. I wouldn't be surprised if Internet-based browsers pop up before long that allow you to compare your genome to others."

The genetic sequence that was unveiled in all its naked detail belongs to J. Craig Venter, the Maryland scientist who led the project at the J. Craig Venter Institute in Rockville, Md. It carries significantly more information than the two previously sequenced human genomes, released in 2001. Those sequences -- one assembled by Venter and co-workers at Celera Genomics and the other by federally funded scientists -- were composites of several people. And although they were referred to as complete, they were in fact half-genomes -- or "haploid" -- containing a parental mosaic of the 3 billion DNA letters that can fit on one set of the 23 chromosomes paired in every cell.

Not emphasized in 2001 was the fact that people have in their cells two versions of each of those 23 chromosomes, one from each parent -- a "diploid" genome. Increasingly, scientists are finding that the difference between being healthy and being sick has to do with how those genomes interact. From Dad, a person may inherit a version of a gene that predisposes to a disease, but from Mom, that person may inherit a protective version -- or an entirely different gene elsewhere on the genome that counteracts Dad's contribution. "I might want to know: Do I have an additive risk from the genomes from both my parents, or did I get some helpful ones from her that counteract the ones from him?" Venter said. Sorting out those details has been daunting. "It's very easy to start mixing up the readouts from each parent because they are so similar," said Samuel Levy, who led the research with Venter. But new sequencing technologies and computational methods allow scientists to chop a person's DNA into pieces and reassemble the maternal and paternal segments independently.

Challenges remain. Although Venter's method produces a 6 billion-letter diploid genome, it does not produce complete paternal and maternal genomes of 3 billion letters each. But it does produce chunks of DNA that are hundreds of thousands of letters long, all from one parent or the other, allowing the most meaningful maternal-paternal comparisons yet. Previous such snippets topped out at about 13,000 letters, too few to be medically informative. And unlike the Human Genome Project, whose focus on individual letters made it blind to many larger mutations or variations involving hundreds or thousands of letters, the newer methods that Venter used capture all sizes. The new work showed, for example, that Venter lacks one parental copy of the gstm1 gene, known to have a role in neutralizing toxins and carcinogens -- perhaps helping to explain why he has had asthma and skin cancer, Levy said.

Venter is not alone in his effort to create personalized sequences, and at least one competitor offered a more tempered view of the work. "I would call this a small, quantitative milestone," said George Church, a Harvard University professor of genetics who is also racing to produce cheap genomes. He said that Venter's sequence, like previous ones, still has many gaps, was cumbersome to make and, at a cost of tens of millions of dollars, was still too expensive.

He also noted that recent research by Scherer had already suggested that human genetic variability is probably at least 0.3 percent, not the 0.1 percent floated in 2001, which Venter uses for comparison. So Venter's new finding of 0.5 percent amounts to something less than a sea change, Church said.