LEADING EDGE

SHAKING THE TREE OF LIFE

Comparative genomics – the study of the genomic sequence of organisms that are related to humans – could ultimately help to identify targets for drug development. **BY JACK MCCAIN,** Contributing Editor

Confucius said that the measure of man is man, but curious creatures may be useful yardsticks in determining the workings of the human body. Careful comparisons of the human genome with the complete genomes of organisms that are closely and distantly related to humans may help researchers pinpoint and ascertain the functional portions of the human genome. This relatively young discipline is called comparative genomics.

In addition to providing insights into disease processes, this method of inquiry is expected, eventually, to identify targets for drug development. Comparative genomics studies of organisms directly involved in human disease (e.g., bacteria, fungi, parasites) could lead to new diagnostic tests and interventions, too. This area is of enormous interest to biologists and paleobiologists as a tool for clarifying obscure relationships among living organisms and the evolutionary paths that they and their extinct predecessors may have followed.

Toward these ends, an international effort is in progress to determine the genomic sequence for numerous organisms selected from branches all over the evolutionary tree (Figures 1–4). Note that the tree's true shape is unknown in many instances and is subject to substantial ongoing revision.

The National Human Genome Research Institute (NHGRI), part of the National Institutes of Health, is coordinating one of the largest sequencing endeavors. Primarily funded by the NHGRI, sequencing of organisms selected is carried out at centers across the country (Agencourt Bioscience, Beverly, Mass.; Baylor College of Medicine Human Genome Sequencing Center, Houston; Broad Institute/MIT Center for Genome Research, Cambridge, Mass.; The Institute for Genomic Research [TIGR], Rockville, Md.; Washington University Medical Center, St. Louis). Organisms selected for sequencing include many with a long history of use as models in biomedical research (mouse, sea urchin) and creatures with agricultural importance (cow, chicken, honeybee).

"We didn't anticipate that every genome would become a project in itself," says Jane Peterson, PhD, associate director of the NHGRI's Division of Extramural Research, who runs the institute's comparative sequencing program.

The resulting genome sequences provide raw material for another ambitious project, ENCODE — the ENCylopedia of DNA Elements. The goal of ENCODE is to compile a complete catalog of the functional elements in the human genome. The majority of these elements are the genes that carry the codes cells use to construct proteins, but some genes encode certain kinds of RNA (ribosomal RNA [rRNA], transfer RNA [tRNA]). Beyond that are the regulatory elements that help determine when various genes are turned on or off.

CONTINUING RESEARCH

In this context, the completion of the Human Genome Project in 2003 (a draft was announced in 2000) was an impressive accomplishment but no more than a beginning. It's essentially a lengthy, ordered list of the basic ingredients needed to assemble an as-yet

unknown number of complex parts.

Specifically, the human genome comprises 3 billion nucleotides from which DNA is constructed. The chemical bases of the four kinds of nucleotides in DNA are two purines, adenine (A) and guanine (G), and two pyrimidines, cytosine (C) and thymine (T). In RNA, thymine is replaced by uracil (U). Per Watson-Crick base-pairing rules, A always pairs with T (or U) and G with C, as occurs in the famous double helix formed by twin strands of DNA. This principle allows complementary strands of DNA or RNA to be assembled from the original template. It applies equally to human cells and bacteria — in fact, to every living thing on the planet that uses nucleic acids to pass along its genetic heritage to the next generation.

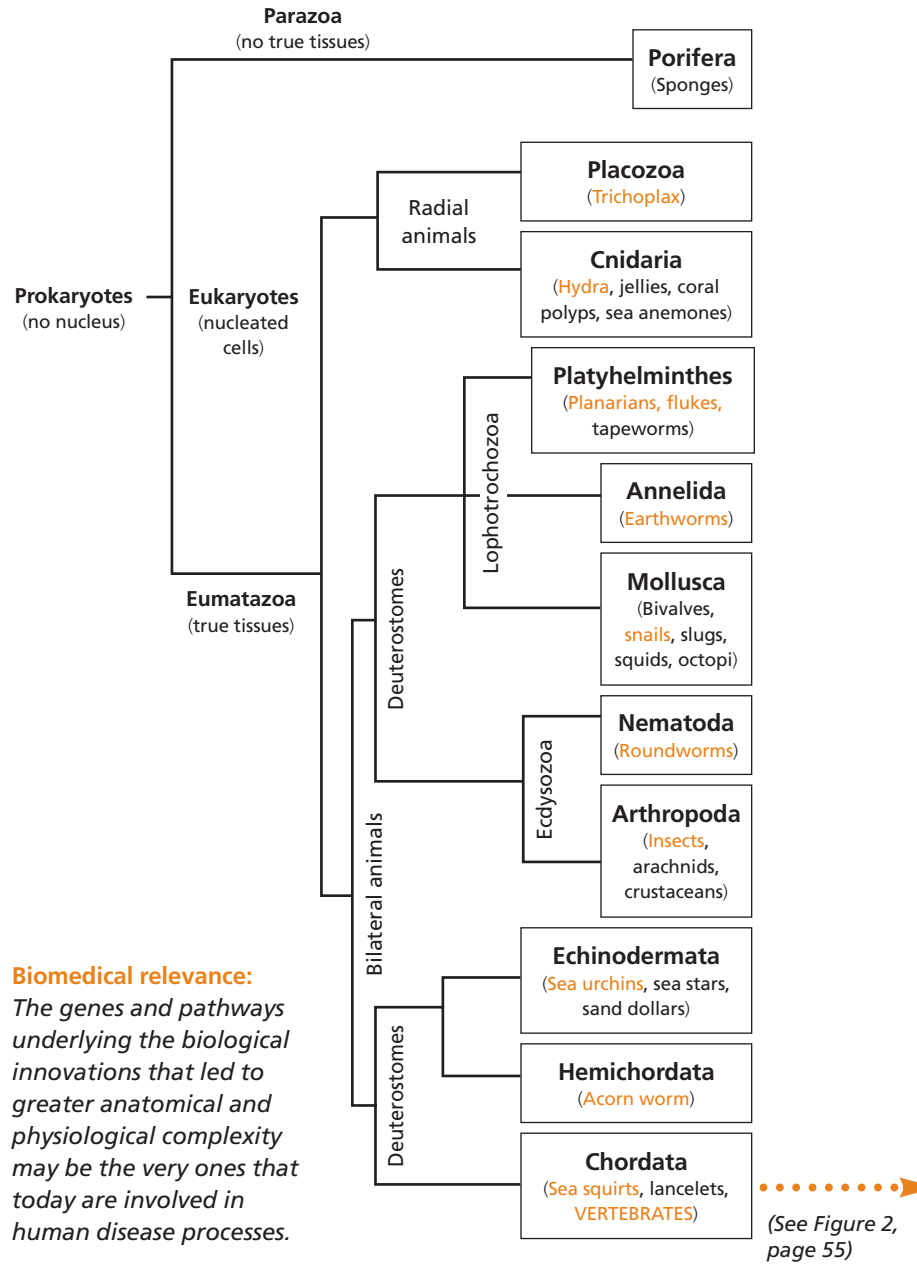
The Human Genome Project has determined the exact sequence in which the 3 billion nucleotides appear. As of October, the overall sequence was regarded as 99.999 percent accurate — just 1 incorrect base per 100,000 base pairs (bp). To find genes and related functional elements, it's important to have long, uninterrupted, high-quality sequences. When the working draft was completed, the length of the average such sequence was 81,500 bp; now, it's 38.5 million bp. A few small gaps remain, but they are beyond the reach of existing technology. Nevertheless, the extant sequence is sufficient for high-quality research.

Assembly of the human genome and other recently completed genomes (e.g., cow, dog, chicken) does not mean the others will just fall into place. "We didn't anticipate that every genome would become a project in itself," says Jane Peterson, PhD, associate director of the NHGRI's Division of Extramural Research, who runs the institute's comparative sequencing program.

She notes that some lower invertebrates present biologically interesting problems in obtaining DNA needed for analysis. Some of the organisms are rare; others have bacteria clinging to them, making it dif-

FIGURE 1 Eukaryotic phylogeny

Organisms whose names appear in color have been selected by the National Human Genome Research Institute for genome sequencing.



Biomedical relevance:
The genes and pathways underlying the biological innovations that led to greater anatomical and physiological complexity may be the very ones that today are involved in human disease processes.

Glossary

Anapsids. Reptiles with no openings behind the eye. Turtles are technically anapsids but may have evolved from *diapsids*.

Base. A nitrogenous ring or double ring at the heart of a *nucleoside* or *nucleotide*. Single-ringed bases are the *pyrimidines* cytosine (C), thymine (T, found in DNA), and uracil (U, found in RNA). Double-ringed bases are the *purines* adenine (A) and guanine (G).

Base pair. A complementary pair of nucleotide bases.

Blastopore. In animal embryos, an opening to the exterior from the primitive gut.

Clade. A set of organisms grouped together on the belief that they evolved from a common ancestor.

Codon. A sequence of three nucleotides that dictates the synthesis of a specific amino acid. With four "letters" available, 64 codes are possible (4x4x4); 61 encode amino acids (most of the 20 amino acids are encoded by more than one codon), and three are stop codes (TAA, TAG, TGA).

Deuterostomes. Organism whose embryonic *blastopore* forms only the future anus; mouth forms later. Echinoderms, hemichordates, and chordates are deuterostomes.

Diapsids. Reptiles with two skull openings behind the eye socket.

Diploblastic. Having two cell layers (ectoderm and endoderm) with mesoglea in between. Most diploblasts have radial symmetry, except for some sea anemones with bilateral symmetry. Compare with *triploblastic*.

DNA. Deoxyribonucleic acid — repository of the genetic code. Found in the nucleus (and mitochondria).

Ecdysozoa. A clade of protostomes including nematodes and arthropods.

Eukaryotes. Organisms whose cells have a nucleus, which contains the DNA, and usually various organelles (e.g., *mitochondria*). During cell division, exact copies of the DNA are made through the process of mitosis. Contrasts with *prokaryotes* — much smaller cells lacking nuclei. A typical eukaryotic cell has about 1000 times the volume of a prokaryote. Fungi, plants, and animals are eukaryotes.

Genome. Complete set of DNA for an organism.

Homologous. Descended from a common ancestor.

Lophotrochozoa. A new clade of protostomes now linked together because of similarities in their genomes. Group includes flatworms, annelids, and mollusks. Compare with *Ecdysozoa*.

Mitochondria. Organelles containing respiratory enzymes, used to generate ATP from food molecules.

mRNA. Messenger RNA — a long, single-stranded molecule of RNA that is transcribed from a gene on DNA.

Nucleoside. *Base* + sugar. In RNA, the sugar is ribose; in DNA, deoxyribose. *RNA nucleosides:* adenosine, guanosine, cytidine, uridine. *DNA nucleosides:* deoxyadenosine, deoxyguanosine, deoxycytidine, deoxythymidine.

Nucleotide. The building blocks of RNA and DNA, composed of a *base* + sugar + phosphate. In the DNA double helix, the bases form the cross-supports, while the sugars and phosphates provide lengthwise structure. *RNA nucleotides:* adenosine monophosphate (AMP), guanosine monophosphate (GMP), cytidine monophosphate (CMP), uridine monophosphate (UMP). *DNA nucleotides:* deoxyadenosine monophosphate (dAMP), deoxyguanosine monophosphate (dGMP), deoxycytidine monophosphate (dCMP), deoxythymidine monophosphate (dTMP or TMP).

Orthologs. Homologous genes that retain the same function as the ancestral gene. Contrast with paralogs — homologous genes that evolve to develop new functions.

Prokaryote. The earliest organisms — small cells lacking a nucleus (and other internal subdivisions, called organelles); DNA drifts loose in the cytoplasm. Prokaryotes reproduce by splitting. All living bacteria and archaea are prokaryotes.

Protostomes. Organisms whose blastopore forms the future mouth and anus. *Lophotrochozoa* and *Ecdysozoa* are protostomes. Compare with *deuterostomes*.

Prototheria. The "first beasts," represented today solely by the monotremes.

Ribosome. The cellular machinery where proteins are assembled, using the codes carried in mRNA.

RNA. Ribonucleic acid. Once thought to serve primarily as a messenger for DNA, carrying DNA's genetic code in the process of gene expression, but now known to play other roles, including enzymatic and regulatory functions.

Therapsids. Mammal-like reptiles.

Theria. Mammalian subclass that includes marsupials and eutherians.

Triploblastic. Having three tissue layers (ectoderm, mesoderm, endoderm). All triploblasts have bilateral symmetry.

tRNA. Transfer RNA — small units of RNA that recognize codons in mRNA.

ficult to separate the desired DNA from bacterial DNA. Top scientists are willing to solve such problems because of the importance of sequence data in providing new biological insights. Sequencing has become essentially a matter of production, albeit a complex process that still has room for improvement in terms of speed and ease of use.

Meanwhile, the fruits of genome sequencing generate a wealth of information destined to hold the interest of computational biologists for decades, as they unravel the intricate relationships among genes, proteins, and regulatory elements.

The human body expresses about 100,000 proteins, but not all at the same time — hence the importance of regulatory elements. A rule of one gene, one protein was once proposed. But when the first draft of the human genome was completed, the estimated number of human genes was revised from 100,000 to 30,000–35,000. As of October, the estimate was reduced to fewer than 25,000, including 19,599 genes known to encode proteins and 2,188 DNA segments thought likely to encode proteins. So, the old rule no longer applies; the average gene now is believed to express three or four proteins.

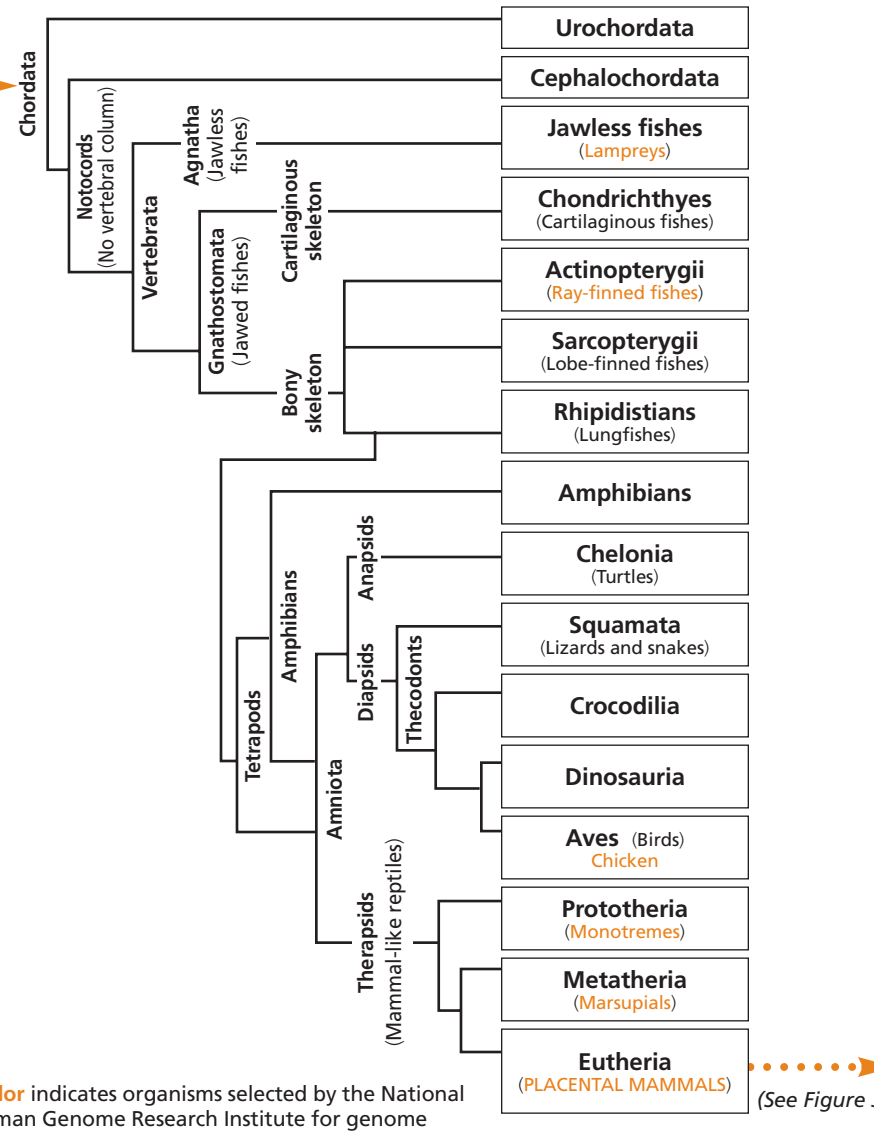
The number of genes does not indicate an organism's complexity. The mustard plant has slightly more genes (26,000), and the fruit fly and roundworm only slightly fewer genes (14,000 and 19,000, respectively) than the human genome.

Genes do not participate directly in protein synthesis. Instead, the code for a given protein is copied (*transcribed*) in the nucleus by messenger RNA (mRNA). Unlike DNA, which is a highly stable molecule, mRNA is transient, surviving a few hours at most. This property allows it to be turned on and off.

A sequence of DNA that includes a gene for a given protein contains the coding that identifies each amino acid in sections known as *exons*. The shuffling of exons and the duplication and subsequent diversification of exons are believed to be important evolutionary mechanisms.

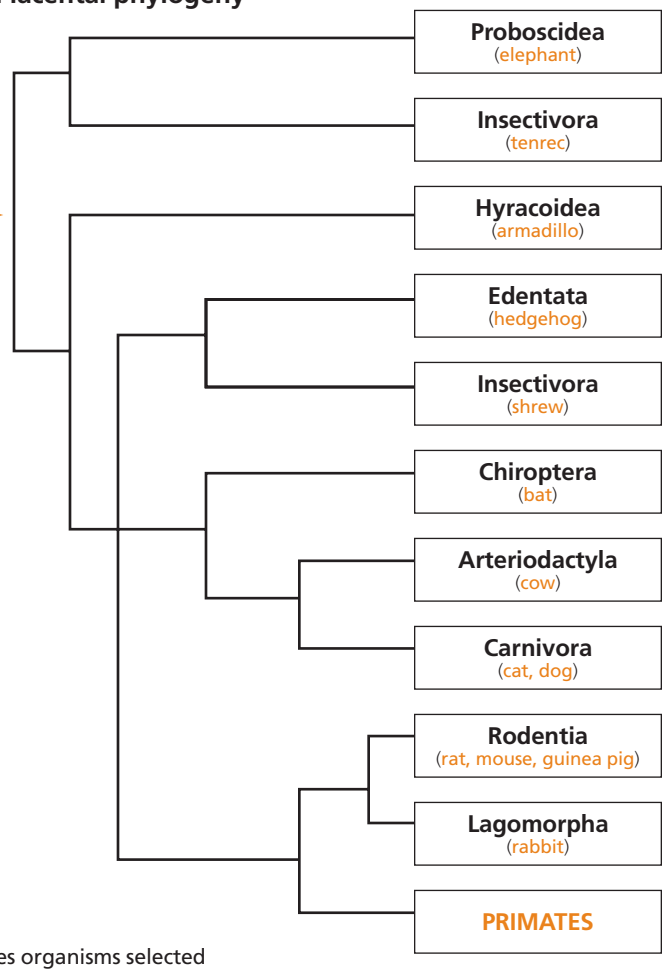
Within a gene, exons are separated by noncoding sections, known as *introns*. After the gene has been transcribed from the DNA, the introns are cast aside and the remaining exons are spliced together to form the finished mRNA molecule. Embedded in the structure of a gene are a specific start signal (the nucleotide sequence ATG) and a stop signal (TAA, TAG, or TGA), which direct the molecular machinery that

FIGURE 2 Vertebrate phylogeny



(Color indicates organisms selected by the National Human Genome Research Institute for genome sequencing.)

FIGURE 3 Placental phylogeny



(Color indicates organisms selected by NHGRI for genome sequencing.)

See Figure 4

transcribes mRNA. Introns also have characteristic sequences marking their beginning (GT) and end (AG).

The mRNA then carries the code from the nucleus into the cytoplasm, where *ribosomes* await to assemble the protein. The code for amino acids is contained in triplets of nucleotides called *codons*. Each of the 20 amino acids is encoded by one or more codons. For example, GGU, GGA, GGC, and GGG all encode glycine. During protein assembly, each amino acid is carried into position along the ribosome by a molecule of tRNA specific for that amino acid.

Exons account for about 1.5 percent of the human genome; introns, 24 percent. The remaining 75 percent is intergenic material. Molecular biologists once regarded these noncoding sequences as junk, but comparative genomics suggests that view may be outmoded. A recent study in *Nature* compared 13 vertebrate genomes (human, chimpanzee, baboon, cat, dog, cow, pig, rat, mouse, chicken, zebra fish, and two species of pufferfish) and found that extensive noncoding sequences of DNA were *conserved* across a wide range of species.

This principle of conservation of DNA sequences down through evolutionary time forms the heart of comparative genomics. The logic says if the same sequences are found in the genomes of organisms as disparate as, say, yeast and humans, these sequences are likely to be homologous (stemming from a common ancestor) and have been conserved due to their fundamental importance to the functioning and survival of the organism. This principle applies equally to DNA sequences that encode genes or to noncoding sequences found amidst the intergenic DNA.

THE LOWLY YEAST

The complete genome of baker's yeast (*S. cerevisiae*) was determined in 1996 by an international consortium involving 92 laboratories. It comprises 12 million nucleotides, divided among 16 chromosomes. Until quite recently, it was thought to contain about 6,200 genes.

The yeast genome has been the object of intense scientific scrutiny, so much so that Manolis Kellis, PhD, a computational biologist at MIT, says, "There seems to be one researcher for every gene."

As a result of so much focused research, concrete functions have been assigned to about 4,000 of the yeast's genes, making this species the most "validated" organism of all. No function has been assigned to the other 2,000 purported genes, however, raising the question of whether they are even real.

Kellis explains that if two genomes contain a common gene consisting of thousands of nucleotides, the probability that the two sequences are the results of random events is low, and the genes stand out amidst the noise. Yet a small pro-

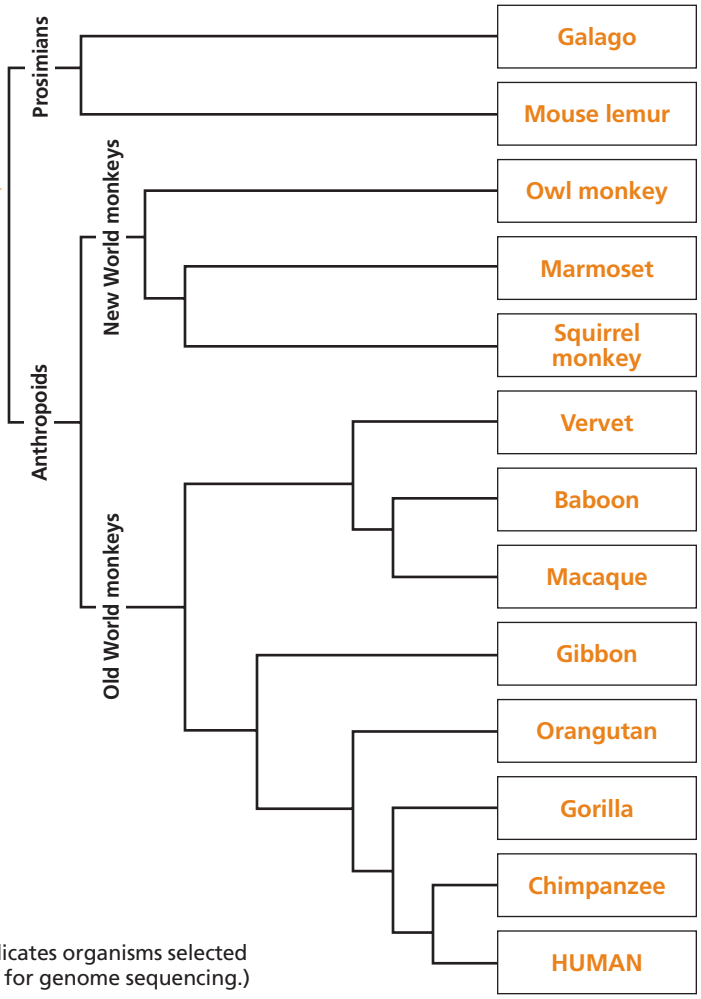
tein consisting of even 20 amino acids would require only 60 nucleotides to transcribe the codons needed to assemble these amino acids. In the absence of computational techniques for detecting such subtle patterns, these small genes are indistinguishable from the abundant noise.

Comparing the *S. cerevisiae* genome with those of three other members of the same genus (*S. paradoxus, S. mikatae,* and *S. bayanus*), Kellis and colleagues developed a technique for identifying subtle but distinctive signatures of evolutionary change. When they applied their technique to the 4,000 genes of known function, 99.9 percent were accepted as valid genes. Likewise, the technique rejected 99.6 percent of the intergenic material. Their technique was highly sensitive *and* specific — capable of providing, for the first time, a definitive yes-or-no answer as to whether a suspected gene really was or was not a gene.

When Kellis and colleagues applied the technique to the other 2,000 suspected genes in the *S. cerevisiae* genome, 1,500 were accepted but 500 were firmly rejected. The study results, published in *Nature,* revealed that, despite extensive previous research, the prevailing picture of the *S. cerevisiae* genome had been inaccurate, with about 15 percent of its purported genes being affected by the team's findings. In addition to rejecting 500 suspected genes, 43 new genes were identified, and the boundaries of more than 300 genes were reset.

Kellis and colleagues are applying this technique to the human, mouse, rat, and dog genomes, separated by about the same evolutionary distance (measured by the number of nucleotide substitutions per site) that separates the four yeast species. Yet the human genome contains much longer stretches of intergenic material and repeated bases, complicating the identification of human genes. The importance of Kellis's technique is to make it easier to distinguish signals from noise in large genomes like those of the human and the mouse.

FIGURE 4 Primate phylogeny



(Color indicates organisms selected by NHGRI for genome sequencing.)

TIME TRAVEL

As seen with the yeast-human comparison, comparative genomics involves traveling back in time, in a limited sense. In comparing the human genome with that of the chimpanzee, humankind's closest relative, one sees genetic changes that occurred in each species in the last 5 or 6 million years, starting when the lineage that led to modern humans diverged from the lineage that led to the great apes. After that interval, only 1.2 percent of the human genome differs from that of the chimp genome, a percentage that undoubtedly holds clues to why humans differ from chimps and for determining why conditions rarely seen in chimps commonly afflict humans

(e.g., rheumatoid arthritis).

But great overall similarity of human and chimp makes a comparison of the two genomes less useful for determining the function of genes in the numerous physiologic processes that the two primates share. Molecular biologists therefore are sequencing the genomes of organisms representing numerous branches on the ancient tree of life.

Today's 4,600 species of mammals descend from those that appeared early in the Mesozoic Era. The most primitive mammals, the Prototheria, are represented today by just one group, the monotremes

TABLE Evolutionary milestones

Millions of years ago	Lineages diverging from common ancestor	Comments (Color indicates organisms whose genomes have been or will be sequenced.)
5–6	Human – chimp	Chimpanzee is closest living relative to human
8	Human – gorilla	Other apes (gorilla, orangutan) nearly as similar to human as is the chimpanzee
12	Human – orangutan	Evolutionary midpoint between humans and Old World monkeys
20–40	Mouse – rat	
25	Human/ape – Old World monkeys	Old World monkeys include rhesus macaque, baboon, pig-tailed macaque, ring-tailed macaque
35–40	Human/ape – New World monkeys	New World monkeys include marmoset, squirrel monkey, owl monkey
60	Anthropoids – prosimians	Lemur (a prosimian) is model for aging research
65	**End of Cretaceous Period and Mesozoic Era** *Extinction of dinosaurs, thought to trigger explosion in mammalian diversity*	
80	Primates – carnivores	
125	Birds – reptiles	
250	**End of Permian Period and Paleozoic Era** *Mass extinction at end of Permian eliminated 83 percent of genera and 57 percent of families, triggering explosion in marine diversity*	
350	Mammalian lineage – bird-reptile lineage	
400	Tetrapods – fish	
520?	Vertebrates – primitive chordates	Surviving representatives of the primitive chordates are the tunicates (sea squirts)
	Protostomes – deuterostomes	Protostomes include Annelida (segmented worms), Mollusca (bivalves, snails, slugs, squids, octopi), and Arthropoda (insects, arachnids, horseshoe crabs, centipedes, millipedes) Deuterostomes include phyla Echinodermata and Chordata
530	Body cavities – no body cavities	Nematoda (roundworms) was first phylum with a quasi-body cavity (pseudocoelom) Platyhelminthes (planarians, flatworms, flukes, tapeworms) are only surviving bilateral animals lacking body cavity
590	**Beginning of Cambrian Period and Paleozoic Era**	
	Bilateral symmetry – radial symmetry	Phylum Cnidaria includes all surviving radially symmetrical animals (hydras, jellies, coral polyps, sea anemones)

Manolis Kellis, PhD, and colleagues have developed a technique for identifying subtle but distinctive signatures of evolutionary change.

—the duckbilled platypus and spiny anteater. The Prototheria diverged into the more familiar Metatheria (marsupials) and Eutheria (placental mammals). Thus the selection by the NHGRI of the platypus, opossum, and wallaby for genome sequencing provides a chance to compare and contrast humans with distantly related mammals.

Because of its position on the evolutionary tree (Figure 2, page 55), the platypus genome will bridge a phylogenic gap between reptiles/birds and eutherians. Comparing the monotreme genome with marsupial and eutherian genomes can help to determine the ancestral mammalian genome.

Combining reptilian and mammalian features with other characteristics unique to itself, the platypus is curious indeed. Discovered 200 years ago in Australia, the first specimen was thought to be a hoax. The platypus lays eggs but secretes milk, and it is the only mammal whose mammary glands lack nipples (the young suck the milk off the surface of the abdomen).

The creature's unique features include a snout equipped with an electrosensory mechanism, which is helpful for finding invertebrate prey, and a venom gland that the males wield during territorial fighting. Although not fatal to humans, the venom results in intense pain, immediate local swelling, and swelling of adjacent lymph nodes. The venom contains unique proteins and peptides that might prove useful in pain management.

Further down the evolutionary tree (Figure 1, page 51), adding a nonmammalian genome to the mix helps identify genes specific to mammals and genes that may have been possessed by an earlier ancestor. Consider this: if you chop a chicken into six pieces, you have the makings of dinner; but if you slice a planarian into multiple pieces, in time you end up with a herd of planarians. Wholesale tissue regeneration is highly relevant to human health, as the human body is relatively limited in its regenerative capacities, notably after spinal cord injury, stroke, and myocardial infarction.

Some relatives of the planarian are parasites that are responsible for worldwide morbidity and mortality (for instance, from schistosomiasis). Through comparative genomics, it may be possible to find genes specific to these platyhelminths and identify targets for pharmaceutical intervention.

Or take another distant relative of humankind—the common honeybee. The hives of hardworking honeybees are hot, humid, and densely packed with thousands of adults and juveniles—a perfect breeding ground for bacteria. The honeybee genome therefore is expected to be useful for efficiently identifying the antibiotics the bees use to fend off pathogens. Thus, these molecules could lead to the development of novel antibiotics for treatment of humans.

Likewise, the bee genome could advance our knowledge of the structure of the proteins in bee venom, which could have therapeutic applications for bee sting allergies, in particular, and allergic reactions, in general. A less obvious application of the honeybee genome is to generate a model for studying mental health, based on bees' complex social interactions.

The NHGRI's Peterson thinks comparative genomics initially will generate few therapeutic insights. "At this point, we're still trying to determine how to use the data and which data are needed," she says.

But the tree of life is heavily laden with fruit, and when it ripens the harvest may be bountiful. BH

For more information about the rationale for sequencing the genomes of organisms mentioned in this article, go to «www.genome.gov/10002154». This page provides a sortable table of the organisms selected by NHGRI. If the name of the organism is underlined, you can follow the link to a page with a link to the PDF of the sequencing proposal ("white paper"). If the organism's name is not underlined, go to «www.genome.gov/Pages/Research/Sequencing/NewGenSeqTargets/Summaries/AHGProposal.pdf» or «www.genome.gov/Pages/Research/Sequencing/NewGenSeqTargets/Summaries/CGEProposal.pdf» for details about these organisms.