

**Genomics & Proteomics**  
Reed Business Information  
Morris Plains, NJ, 07950

[Email the editor](#)

[E-mail to a colleague](#)

[Printer Friendly Format](#)

## Cover Story

### All in the Family

**Comparative genomics sheds new light on vertebrate evolution and tells a tale about noncoding regions**

**Vivien Marx**  
Senior Editor



**He was a little right. Carolus Linnaeus (1701-1778), natural philosopher and religious man, so-called father of taxonomy, who invented the system of naming species was not off the mark. Today's genomic insight sheds new light on his ideas of classification based on similarities and the close relations of all primates. (Source: Hunt Institute for Botanical Information Carnegie Mellon University)**

"fixated on genes," finding them and counting them. "I think that is all very important but, by the way, it is only a minority of the functionally important part of the human genome," he says. So now genomics and bioinformatics researchers have their work cut out for them. There is also plenty of room for tool development in the data-crunching intense analysis involved in cross-species comparisons. What follows is a snapshot of some current work in the field.

#### Humans on top?

The emerging data is steadily revealing that humans are not the pinnacle of evolution. Morris Goodman, Derek Wildman and colleagues at Wayne State University School of Medicine stated in a paper published in the *Proceedings of the National Academy of Sciences U.S.A.* "The accumulating DNA evidence provides an objective non-anthropocentric view of the place of humans in evolution. We humans appear as only slightly remodeled chimpanzee-like apes."

Such concepts are not just short lessons in humility. Close connections between vertebrates are important also because a distinct path connects evolutionary insight and biomedicine. As evolutionary biochemist Wilfried de Jong of the University of Nijmegen, Nijmegen, Netherlands points out, rodents are the favored model animals for human diseases. "It therefore is essential to know how closely or distantly related, in evolutionary terms, they are to humans," he says. After all, if dogs or cows were more closely related to man, which they are not, those animals might be better models, he says.

Multi-species comparison of genomic sequences is an important part of ENCODE (Encyclopedia of DNA Elements), the new three-year, \$36 million consortium coordinated by NHGRI, and which focuses on gene function. The first grants have just been awarded to academic and corporate endeavors, for example the Affymetrix GeneChip technology, to figure out how to gear existing technologies for the large-scale quest of functional elements. "To really understand what all the

"He is a rat." "Stop monkeying around." Slang makes animals out to be more nasty than nice. This linguistic usage might well be part of the longstanding effort to distance humans from the rest of the animal kingdom. Given the dynamic developments in vertebrate genomics, however, the metaphors lose some of their impact. Genomic researchers in this field are partnering new ideas with sophisticated bioinformatics instruments to kick humans right where it hurts: in their superiority complex. Guiding these insights is evolution, or what can perhaps be termed the ultimate life science tool.

The data emerging from comparative analyses of vertebrate genomes is rearranging the phylogenetic tree and kickstarting the field of functional genomics. Looking at family relations in this genomic sense is helping researchers move beyond the phase of identifying and sequencing genes to understanding the dynamics of the genome and functionally analyze coding as well as noncoding genomic regions; to essentially show where the noise ends and the signal begins.

Francis Collins, director of the National Human Genome Research Institute (NHGRI), Bethesda, Md., has called it a way of looking into "evolution's lab notebook" and seeing what has worked. Obtaining genomes of various species and lining them up in ever bigger ways shines a spotlight on the conserved regions of correspondence with unknown function. As Eric Green, scientific director of the NHGRI, says, we have been



elements of the genome mean, which are important and which are not, you need comparative sequence data," says Goodman.

### Expanding the genus *Homo*

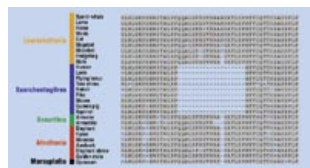
While humans and mice share a common ancestor around 85 million years ago and primates started to diverge around 75 million years ago, when the lineage to lemurs and galagos separated from that of higher primates, humans and their closest living relatives, the common and bonobo chimpanzees, probably only parted ways around 5 or 6 million years ago. In actuality, the numbers are not quite this straightforward and the evolutionary history of mammals is mired in controversy. One debate involves the rate of evolutionary change in different lineages. Sudhir Kumar, PhD, a biologist at Arizona State University, Tucson, differentiates between the rate of mutation and the evolutionary rate of change, in other words the rate with which the change is fixed in the genome. "Mouse and rat may be funky in the sense that they mutate faster somehow than primates, but one shouldn't draw conclusions on all rodents from just one or two species, we need more comparison than that," Kumar says.

There is great disagreement about when the various branching events did occur, and one big issue is that many paleontologists are reluctant to accept the molecular tree (see figure, this page) and to reconcile the molecular tree with paleontological evidence. As lore has it, a bus at a conference was transporting scientists across New York City. Two scientists were heatedly debating phylogenetic relationships of different species. Then, gunfire broke out on a street through which the bus was driving. Everyone in the bus hit the floor, except for the two researchers who kept discussing through the disturbance.

### In the primate bloodstream

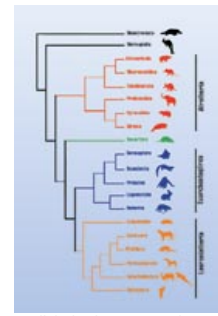
Despite the disagreements, it is clear that chimps and humans are very close relatives. That similarity means it is both quite tricky but also scientifically and medically of great interest to compare their genomes. With a technique they developed called phylogenetic shadowing, Edward Rubin, MD, and his colleagues at the US Department of Energy's Joint Genome Institute and the Lawrence Berkeley Laboratory are sequencing specific segments of primate genomes to hone in on those very small differences as they study functional aspects of genes only found in primates and which appear to indicate heart disease risk. Rather than compare distantly related species, Rubin and his group chose precisely the small genomic differences between non-human primates, combined these differences to create a "shadow", which was then compared to the human genome. With this method the researchers found both exons and introns for several genes as well as the regulatory elements for apo(a). Apo(a), which is associated with low-density lipoproteins and indicative of cardiovascular risk, is found in humans, apes, and Old World monkeys but seems not to be present in other mammals.

Other researchers choose a broader focus. For one study, Morris Goodman, PhD, recipient of the Charles R. Darwin Award for Lifetime Achievement, along with Derek Wildman, PhD, and colleagues used clustal algorithm MACVECTOR 7.0, to compare approximately 90 kb of coding DNA sequence from 97 human genes with their sequenced chimpanzee as well as available gorilla, orangutan and Old World monkey counterparts. In viewing these coding DNA regions, they studied both nonsynonymous substitutions, which are amino acid-changing and thus functionally important, as well as synonymous substitutions which are amino acid-unchanging and thus functionally much less important. The phylogenies, distances, and reconstruction of ancestral sequences were conducted using PAUP\* and MEGA2. The team applied the methods maximum parsimony and maximum likelihood for phylogeny reconstruction and the ancestral sequences were reconstructed with the delayed transformation, or DELTRAN, algorithm. A variety of distances between aligned sequences were calculated using PAUP\* and MEGA2.



click the image to enlarge

**Find the gap. In order to deduce phylogenetic information from multisequence comparisons, scientists search for rare genomic changes such as indels (or insertion/deletions) and transposable elements that are unique to a certain clade. Deletions in the SCA1 protein and other evidence support**



click the image to enlarge

**Molecular tree DNA evidence, phylogeny reconstruction methods and models deliver a topology, or branching pattern of mammalian phylogeny. The dotted lines indicate aspects for which further corroboration is needed, and there is some debate about the molecular dating of the various branching events. The orders Afrosoricida, Eulipotyphla and Cetartiodactyla are recent orders developed due to genomic evidence. (Source: Adapted from Murphy *et al.* 2001, courtesy of Ole Madsen and Wilfried W. de Jong)**

The scientists looked, for example, at four genes that encode for cell receptor proteins and which are likely to be under strong selective pressure. The rate of nonsynonymous change decreases by approximately 12-fold on the human- and chimpanzee-terminal branches of the phylogenetic tree compared with the stem of the tree. Chimps and humans share 99.4% of the sequences at nonsynonymous sites and 98.4% at synonymous sites. So, nonsynonymous substitutions in this group of positively selected genes show that chimpanzees diverge from the common ancestor about as much as humans do. The scientists assert that such functional DNA data lends support to the notion that humans are not a diverging lineage outside the ape clade. Instead, it appears gorillas diverged from the common lineage about six to seven million years ago and then humans and chimpanzees only split up around 5 or 6 million years ago.

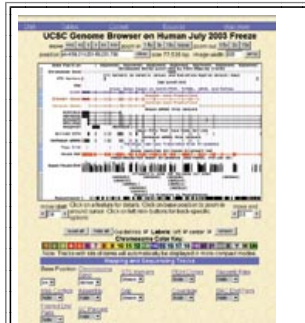
Goodman sparked much debate when, in 1962, he first proposed the idea that chimpanzees and gorillas are more closely related to humans than to other apes and belong in the family

**Euarchontoglires. [Figure adapted from C. Poux et al., *Molecular Biology and Evolution*, vol. 19, pp. 2035-2037 (2002)]**

bonobo chimpanzee.

"This idea that we are so different from other animals, I think, is part of our human chauvinism," says Goodman. To a certain extent, modern vertebrate genomics is proving Carolus Linnaeus (1701-1778), the originator of the binomial system of naming, ranking and classifying all beings, to be right. Without evolutionary concepts but with a rational approach to classification, he placed chimpanzees in the genus *Homo* as *Homo troglodytes*. "He could not see anything exceptional about humans and that they should be widely separated from other creatures," says Goodman. Molecular biologists have always been emphasizing that the basic features of life are held in common by all living creatures. Work on the Hox genes began to help show the extent of genetic correspondence, Goodman says, and that can now be expanded by looking at whole genomes. "I think we should take advantage of having genomes completely sequenced to more thoroughly evaluate the different types of DNA you find in the genome and what they are up to and how they may have shaped our present genetic constitution," he says.

Hominidae rather than Pongidae. As molecular evidence comes in and the ability to analyze it increases, acceptance for this view is growing. The family Hominidae should, in his view, include all existing apes, and the Genus *Homo* should include : Homo (*Homo sapiens* or humankind, *Homo (Pan) troglodytes* or the common chimpanzee and *Homo (Pan) paniscus* or the



**Browsing the Genome**

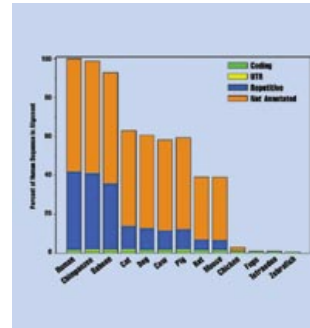
"We have a very big following: we have 140,000 page requests per day and we estimate we have about 4,000 distinct users daily," says Howard Hughes researcher David Haussler, PhD, at the Department of Computer Science at the University of California in Santa Cruz (UCSC).

There are basically three information portals from which to obtain and analyze large genomic datasets: one at the National Center for Biotechnology Information (NCBI), another, Ensembl at the Wellcome Trust Sanger Institute and then there is the Human Genome Browser developed by Haussler and his colleagues. "It is the one that has, in my biased view since David is a collaborator, embraced what I think is going to be the new world of multispecies sequence data availability and is about building the kinds of tools we need to do that," says Eric Green.

The Human Genome Browser began when Jim Kent, then a PhD student whom Haussler was advising, developed the algorithm GigAssembler, to arrange a mess of 400,000 sequence fragments into what became the draft of the complete human genome. "So the browser started out as a

**Joining Mice and Men**

If looks, or morphology, count for much, then primates and rodents are not closely related. In the past, as evolutionary biochemist Wilfried de Jong at the University of Nijmegen says, mitochondrial DNA was the "golden standard" for phylogenetic study because it is easier to isolate, it contains few genes, and meiosis is not involved in its inheritance. "Mitochondrial sequences strongly separate rodents from primates," de Jong says. The analysis of concatenated mitochondrial DNA, however, leads to different conclusions than the analysis of individual genes. And so, de Jong says, particularly when traveling into deeper phylogeny for comparison, with deep meaning more than 40 million years, mitochondrial DNA analysis delivers contradictory results. The remedy for that issue may lie in the nucleus.



click the image to enlarge

[Adapted from J.W. Thomas et al., *Nature*, vol. 424, pp. 788-793 (August 14, 2003)]

With comparisons of nuclear genes becoming more feasible, the fog over the primate-rodent relationship is lifting. De Jong and his colleagues were studying two completely independent genes involved in neurodegenerative disorders and two deletions caught their eye. One is a large deletion in exon 8 of the gene for spinocerebellar ataxia (SCA1) resulting in a deletion in the encoded protein and the other is a 6 base-pair deletion in the prion protein gene (PRNP). They sampled all mammalian orders, looked at the comparative alignments of these genes on their different chromosomes. As it turns out, both deletions are common to all the Euarchontoglires, in other words the orders Primates, Rodentia, Lagomorpha (rabbits), Scandentia (tree shrews) and Dermoptera (flying lemurs).

This indel, or insertion/deletion, is absent in the clades *Xenarthra* (includes armadillos), *Afrotheria* (includes elephants), *Laurasiatheria* (includes the carnivores), and the marsupials, which are an outgroup since they are nonplacental mammals. Adding this closest relative to placental mammals into the mix contributes to phylogenetic analysis as well as the understanding of the relationships between placental mammals. In summing up the picture thus far, de Jong says, "Originally, when the marsupials and the placental mammals diverged, the deletion did not yet exist," he says. "It occurred in the last common ancestral lineage of Euarchontoglires."

The two independent genes under study show a very rare deletion for the same group of animals. "So that means you have a rare event occurring twice in the same group, which makes it much more significant," he says. The study of indels thus supports the view that rodents and primates belong in the same superordinal clade of placental mammals.

Web tool and has evolved into an elaborate one," says Haussler. Kent, the primary designer, has 10 years of experience in the software industry as well as a PhD in molecular biology. "He has brought in two different perspectives, an understanding of what molecular biology researchers need in a browser and common sense about a Web-based tool," says Haussler.

The UCSC team has grown to 30 people who develop the algorithms along with external collaborators, creating what Haussler calls a unified view of the vertebrate genome. "The view allows you to cruise along the chromosomes, zoom in and zoom out, and you can select the features you want to see," he says. The Browser is an integrator which links as various tracks the different available methods and datasets for analyzing the genome. Exploring a region with this bioinformatics tool means being able to compare and contrast the varied available information about it.

The browser is written in C, and all of the Web interfaces are done in simple CGI. That was a conscious decision because Java and Flash are not as ubiquitous as CGI, says Haussler. "The point is, this will run on grandma's old Mac," he says. "When you are talking with a diverse audience of molecular biologists you need to have something reliable and that works on all platforms."

UCSC uses the same reference sequence as NCBI and Ensembl. These portals are all interlinked and use the same freezes and updates. "It is very important that we are talking about the same sequence," he says. Agreeing on the genes themselves, he says, would be advantageous but that is an area where concepts and methods are still emerging. And then there are plenty of genomic regions in which the gene structure is not clear at all.

For Haussler, bioinformatics for comparative genomics starts with blastz, a program authored by computational biologist Webb Miller at Pennsylvania State University, and which matches sequences at the base level. Individual bases can be aligned across species. "It is wonderful we can do this in mammals. They are not so diverged from each



### Finding signposts

There are different phylogeny reconstruction methods, models and algorithms to compare different genes, and de Jong cautions that it is possible to obtain different results with the same dataset. And actually some scientists, like Sudhir Kumar, do not position rodents and primates as such close relatives. Whichever their working hypothesis, in order to have confidence in a given clade, de Jong says that congruence between topologies obtained from independent molecular

data sets and analyses is of the utmost importance. "The most convincing manner to deduce phylogenetic information from multisequence comparisons is to search for rare genomic changes, such as indels and transposable elements, that are unique for a certain clade," he says.

Indels in protein-coding DNA are evidence of a complex mutational mechanism, a more complex one than a single base substitution. Because of this complexity, indels function as phylogenetic signposts. Finding this kind of signage is the next big challenge in vertebrate genomics. Eric Green, scientific director of the NHGRI and director of the NIH Intramural Sequencing Center, along with other researchers are setting out to do just that.

### When it doesn't code

In recently published work [*Nature*, vol. 424, pp. 788-792 (14 August 2003)], Green and colleagues found three insertions that were not previously known and which, in their view, support the rodent-primate grouping. A key facet of this work is that it involved analyzing large blocks of vertebrate sequence: 12 megabases of sequence from 12 species all derived from the genomic region orthologous to about 1.8 megabases on human chromosome 7 which contains 10 genes, including the one mutated in cystic fibrosis. This study, "represents the most diverse collection of large blocks of orthologous vertebrate sequence generated to date," say the authors.

"Nobody had genomic sequence data like that," Eric Green says. "But when you lay out a megabase stretch, you can see genomic events that took place in this case, transposon events and you can figure out at what stage the splitting up of different lineages happened." When you have complete sequences, he says, anecdotal evidence based on snippets of sequence pales in comparison.

With this study, Eric Green, along with Webb Miller and his colleagues at Pennsylvania State University, Philip Green and colleagues the University of Washington and David Haussler and others at the University of California at Santa Cruz (UCSC) along with colleagues at the Children's Hospital Oakland Research Institute and the New York State Department of Health have delivered further evidence for a non-anthropocentric universe. In addition, as they write, they were also able to identify "substantial numbers of conserved noncoding segments beyond those previously identified experimentally, most of which are not detectable by pair-wise sequence comparisons alone." The function of these regions is not clear. Their existence is evolution's way of pointing out that they are indeed of great significance.

### Being naïve is a start

While satisfied with the progress made to date in genomics, Eric Green does not hold back with a call for modesty: "We have been obsessed with genes, finding them, and all of that . . . in reality, I think we will look back on these days of sweat and hard work . . . and we will say 'boy was that easy' because we are so, so, so naïve about the remaining functional part of the human genome. That's the noncoding part." One major challenge, Green says, is that "we simply do not know what we are looking for." While there are rules that govern what a gene sequence should be, such as codons, start, and stop sequences as well as computer tools to find them, the genes, says Green, are going to turn out to be only a minority of the functionally important part of the human genome. Mouse-human comparisons reveal that 5% of the human genome

other that the history isn't implicit in the sequence itself," he says.

Some genomic regions are trickier, for example, due to rearrangements. Or they may deliver surprises, for example when noncoding regions align well across species.

is functionally important, and one-third of that sequence is composed of protein-coding genes. "We have been so fixated on genes and now we are waking up to realize that genes are going to be less than half the story, maybe only a third of the story," he says.

"Why should anyone be interested in a lot of weird organisms, and comparing sequences?" Eric Green asks. The point is not to find genes. "It is to try to start putting down our highlighter, if you will, on parts of the genome and highlight them as being highly conserved across many different species, which provides us a clue that they must be functionally important. Now, we can figure out what they do." The reason, he says, that comparative sequence analysis is so powerful is because "we have so little else." In his view, "we don't know what we are looking for and we don't have a lot of tools." For genes, there are cDNA libraries. No such libraries exist for noncoding functional DNA.

To identify the regions of sequence conservation in their study, the team used blastz in order to construct the sequence alignments and MultiPipMaker to compute the alignments of similar regions. The Human Genome Browser at the UCSC was both the main visualization and dissemination tool for this ongoing work. (see sidebar story, p. 22) "This data gets pretty complicated," Green says and points out that it is one thing to have the human genome sequence and another to compare it to many species and analyze the data in several different ways. "The Human Genome Browser is wonderful," says Green. "It is a functional access point to the data and the analyses of that data are all done in a very visual way."

Pipmaker output is delivered as "percent identity plots," which gives a view of the relationships of more than two sequences. The scientists found that alignments between human and the compared animal sequences revealed the evolutionary relationships in the coding exons as well as noncoding regions. And the analysis of transposable element insertions highlights the variation in genome dynamics between species.

#### Sequences on the move

In the past, Roy Britten of the California Institute of Technology, Pasadena, Calif., and other researchers have pointed out that transposable elements are a major source of mutation and can affect gene expression as they often carry regulatory elements. They are thus key contributors to evolutionary variation. Because of such genomic features, as Eric Green explains, molecular evolution studies need to be multifaceted. This recent comparative analysis of a large stretch of sequence by Green and his colleagues, which looked at exonic changes, neutral substitutions, and transposon events, provides a more robust and informative phylogenetic analysis than just the focus on a single type of genomic change. This large data bundle of evolutionarily diverse genomic sequence is just the beginning, he says. The collection is growing, covering more than 30 species to date.



This focus on targeted regions of the genome in various animals is offering a novel view through the window of vertebrate genome evolution. Depending on the time of day, a window can provide either a clear view or a reflection. Comparative sequence analysis is similar in that respect. The evolutionary insight it delivers reflects back new questions about functional regions of both human and animal genomes.

#### Organizations mentioned in this article:

[Eric Green](#)

[Webb Miller, Penn State Bioinformatics Group](#)

[PipTools: a collection of programs for PipMaker input and processing PipMaker output](#)

[Edward Rubin, director of the US Department of Energy's Joint Genome Institute \(JGI\) and also Director of the Genomics Division at Lawrence Berkeley National Laboratory \(LBNL\).](#)

[Morris Goodman, Wayne state University School of Medicine](#)