Cornell University
Center for Advanced Computing

CAC
we enable your success

CAC and Carlos Bustamante, Adam Siepel, and Andrew Clark

Macaque Genome Analysis: Improving the Ability to Identify Human Genes involved in Diseases

How does the genome of the rhesus macaque differ from human beings?

Finding the Answer

After the macaque genome was sequenced, a scientific team from Cornell was recruited to analyze the results. The Cornell analysis was performed by research groups under Adam Siepel and Carlos Bustamante, Assistant Professors of Biological Statistics and Computational Biology, with assistance from Andrew Clark, Professor of Molecular Biology and Genetics. Richard Gibbs of Baylor College of Medicine oversaw the entire project.

Understanding the Macaque Genome

The rhesus macaque is physiologically similar to humans and therefore widely used in medical research, particularly in vaccine testing and as a model for AIDS research.

Understanding its genome and how it differs from that of human beings promises to offer new insights into the evolution of humans and other primates and has important implications for medical research.

Photo Credit: The Humane Society

Improved Research

Research Metrics

- Speed: Decrease compute time for modeling genome evolution with Cornell’s Computational Biology Service Unit and CAC high-performance computing systems.

Research Challenge

The rhesus macaque genome consists of 2.9 billion DNA base pairs. The key to analyzing much of this data is the availability of powerful computational methods for modeling genome evolution. There are several challenges in this area ranging from bioinformatics and annotation of the gene sequences to the numerical optimization of complex algorithms.

Solution

Cornell researchers used a dedicated computational biology cluster at CAC for genome analysis. An article in the journal *Science* by the Rhesus Macaque Genome Sequencing and Analysis Consortium reported on evolutionary and biomedical insights. A Bill Steele *Cornell Chronicle* article details the Cornell team's contribution:

"Siepel's group studies genes that were found to be common to humans, macaques and chimpanzees. They identified 10,376 genes whose function is at least partially known, and looked for differences that would show how evolution had progressed. By comparing genes that have had 25 million years to change (as compared to the 6 million year gap between humans and chimpanzees), the researchers can learn something about how and why those changes took place.

Over time, minor changes in genes occur randomly, often without changing the amino acids – protein building blocks – for which the genes encode. Siepel's group used these changes as an indicator of how much random change should be expected over 25 years. Then they looked at changes that would code for a different amino acid, which might cause a change in function, and compared these with the expected rate of change.

'Where the amino acids have changed more than you'd expect it's possible nature has responded to some environmental effect,' Siepel explains. For example, the researchers found the most evidence for positive selection in a gene coding for keratin, a protein involved in the formation of hair shafts. Perhaps humans are less hairy than monkeys because of an ancient climate change or some shift in the standards of mate selection, the researchers speculate. Other genes that seem to have been selected for over the years include several involved in the immune system and cell-membrane signaling systems.
On average, the researchers say, genes in the human and chimpanzee genomes have evolved more rapidly than in the other primates, after adjusting for random rates of change.

Siepel's group also analyzed genes that are duplicated in several different locations on the genome. They zeroed in on a family of genes known as PRAME (preferentially expressed antigen of melanoma) that are active in cancer cells and seem to be involved in the formation of sperm. Humans have at least 26 copies. Comparison with the mouse genome suggests that there was a spurt of duplication of this gene early in primate evolution, and comparison with the macaque shows another spurt of copying in both humans and chimpanzees, with the greatest duplication in humans and with evidence for positive selection. This suggests, the researchers say, that the PRAME family has played an important role in human evolution.

Bustamante’s group, who used CAC HPC systems to enable their research, studied variations within the macaque genome – the ways in which individuals within the species differ from one another. While the complete genome sequencing of the macaque was done with the DNA of a single individual, for studies of variation researchers at Baylor also sequenced part of the genomes of 16 other macaques, eight from China and eight from India, and targeted five regions of the genome for deeper analysis, sequencing those regions in fine detail in 47 individuals. Macaques show less variation on the X chromosome (one of the two sex chromosomes) than on others, Bustamante’s group found.

‘Evolutionary theory predicts that if natural selection is important in shaping the sex chromosome, there will be less variation on the X,’ Bustamante says. Since males have only one X chromosome, he explains, a change can’t hide on the recessive side for a few generations and escape selection pressure. A surprise finding was that variation in the X chromosome was only 50 percent of what was seen on the other chromosomes, whereas about 75 percent had been expected.

The researchers also saw substantial differences between the Indian and Chinese macaque populations, which they said could be due to sweeps of natural selection or major differences the histories of the two populations.

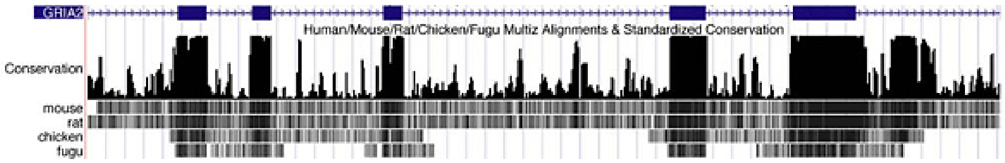
Ryan Hernandez, a graduate student in Bustamante’s group, led an analysis of the difference between Chinese and Indian macaques as well as variations in each of those populations. The analysis suggests that the two populations separated about 162,000 years ago. Both Indian and Chinese macaques are used in biomedical research, and understanding the genetic difference between the two populations is important, Hernandez says. For example, he points out, the simian immunodeficiency virus (SIV) is used as a model for the human immunodeficiency virus (HIV), but when exposed, Chinese macaques develop AIDS-like symptoms more slowly than Indian macaques.

An important finding for medical research, Hernandez says, is that you can travel much faster along the DNA strand in the Indian macaque than in the Chinese macaque before finding a difference between individuals. Researchers looking for a disease-causing gene don’t usually find the exact DNA sequence of the gene right away. Instead, they first determine that the gene is somewhere between two easily recognized sequences called markers and zero in from there. In Indian macaques, Hernandez says, those markers can be farther apart, making the search easier. From there, he suggests, the search could be continued with Chinese macaques, using markers closer together. It is often easier to track a gene in a controlled population of laboratory monkeys than in humans, but since the two genomes are so similar, once it is found in the macaque it can usually be located in humans.

The work on variations, Bustamante said, will help in the development of a dense genetic map for macaques that will ultimately improve scientists’ ability to identify human genes involved in such diseases as cancer, diabetes, and heart disease.”

The Collaborative Relationship

CAC HPC systems and staff at the Cornell CBSU help to enable life sciences research at Cornell.



Provided: Siepel

“The CAC cluster has been integral to the development and testing of our computationally intensive methods. For example, a program can take several hours to analyze a typical data set on a single processor. Since many data sets need to be generated and analyzed to confirm the analysis, we require several thousand CPU hours to validate the method. Without CAC resources, such validation would not be possible. The Computational Biology Service Unit (CBSU) staff has helped us to parallelize our algorithms, reducing our analysis time for a single data set by an order of magnitude and has helped us make our methods available to the rest of the community by providing a Web interface and servers.”

Carlos Bustamante
Assistant Professor of Biological Statistics and Computational Biology
Cornell University