

NOVEMBER 23, 2006

Genetic Variation: We're More Different Than We Thought

New research shows that at least 10 percent of genes in the human population can vary in the number of copies of DNA sequences they contain—a finding that alters current thinking that the DNA of any two humans is 99.9 percent similar in content and identity.

"It may be that some diseases are caused by copy number variation rather than SNPs."

- Stephen W. Scherer

This discovery of the extent of genetic variation, by Howard Hughes Medical Institute (HHMI) international research scholar Stephen W. Scherer, and colleagues, is expected to change the way researchers think about genetic diseases and human evolution.

Genes usually occur in two copies, one inherited from each parent. Scherer and colleagues found approximately 2,900 genes—more than 10 percent of the genes in the human genome—with variations in the number of copies of specific DNA segments. These differences in copy number can influence gene activity and ultimately an organism's function. To get a better picture of exactly how important this type of variation is for human evolution and disease, Scherer's team compared DNA from 270 people with Asian, African, or European ancestry that had been compiled in the HapMap collection and previously used to map the single nucleotide changes in the human genome. Scherer's team mapped the number of duplicated or deleted genes, which they call copy number variations (CNVs). They reported their findings in the November 23, 2006, issue of the journal *Nature*.

Scherer, a geneticist at the Hospital for Sick Children and the University of Toronto, and colleagues searched for CNVs using microarray-based genome

scanning techniques capable of finding changes at least 1,000 bases (nucleotides) long. A base, or nucleotide, is the fundamental building block of DNA. They found an average of 70 CNVs averaging 250,000 nucleotides in size in each DNA sample. In all, the group identified 1,447 different CNVs that collectively covered about 12 percent of the human genome and six to 19 percent of any given chromosome—far more widespread than previously thought.

Not only were the changes common, they also were large. "We'd find missing pieces of DNA, some a million or so nucleotides long," Scherer said. "We used to think that if you had big changes like this, then they must be involved in disease. But we are showing that we can all have these changes."

The group found nearly 16 percent of known disease-related genes in the CNVs, including genes involved in rare genetic disorders such as DiGeorge, Angelman, Williams-Beuren, and Prader-Willi syndromes, as well as those linked with schizophrenia, cataracts, spinal muscular atrophy, and atherosclerosis.

In related research published November 23, 2006, in an advance online publication in *Nature Genetics*, Scherer and colleagues also compared the two human genome maps—one assembled by Celera Genomics, Inc., and one from the public Human Genome Project. They found thousands of differences.

"Other people have [compared the two human genome sequences]," Scherer said, "but they found so many differences that they mostly attributed the results to error. They couldn't believe the alterations they found might be variants between the sources of DNA being analyzed."

A lot of the differences are indeed real, and they raise a red flag, he said.

Personalized genome sequencing—for individualized diagnosis, treatment, and prevention of disease—is not far off, Scherer pointed out. "The idea [behind comparing the human genome sequences] was to come up with a good understanding of what we're going to get when we do [personalized sequencing]," he explained. "This paper helps us think about how complex it will be."

In a "News and Views" article in the same issue of *Nature*, HHMI professor Huntington F. Willard writes, "the stage is set for global studies to explore anew...the clinical significance of human variation." Willard is director of the Institute for Genome and Science Policy at Duke University.

To fully extract meaningful data using the human genome maps, researchers must know what's missing and how much variation exists, Scherer said. "Our computer algorithms are smart, but it is hard to find something if it is not there in the reference you are comparing against."

In fact, Scherer's group found some 30 million nucleotides that are seemingly not yet represented at all, or in different copy numbers or orientations, when comparing the Celera assembly to the public human genome sequence. The entire human genome is thought to contain about 3 billion nucleotides.

The discovery of an abundance of DNA variation puts a whole new spin on the study of genetic disease. Most research has focused on small alterations, called single nucleotide polymorphisms (SNPs). It may be, said Scherer, that some diseases are caused by copy number variations rather than SNPs. In fact, recent research has already linked such variations to kidney disease, Parkinson's disease, Alzheimer's disease, and AIDS susceptibility.

The discovery also provides a new outlook on evolution.

"Until now, our focus has been on examining evolution through either small SNP changes or larger chromosomal alterations you can see under the microscope, because that's what we could detect," Scherer said. "But now there's a whole new class of mid-sized variants encompassing millions of nucleotides of DNA to consider."

This change in the way scientists think about human genetics is exciting, but it is still very early to know what all this means, said Scherer. "Though it does make you wonder, he added. "If you have 1 million fewer nucleotides than your buddy, shouldn't you get a break on your golf handicap?"