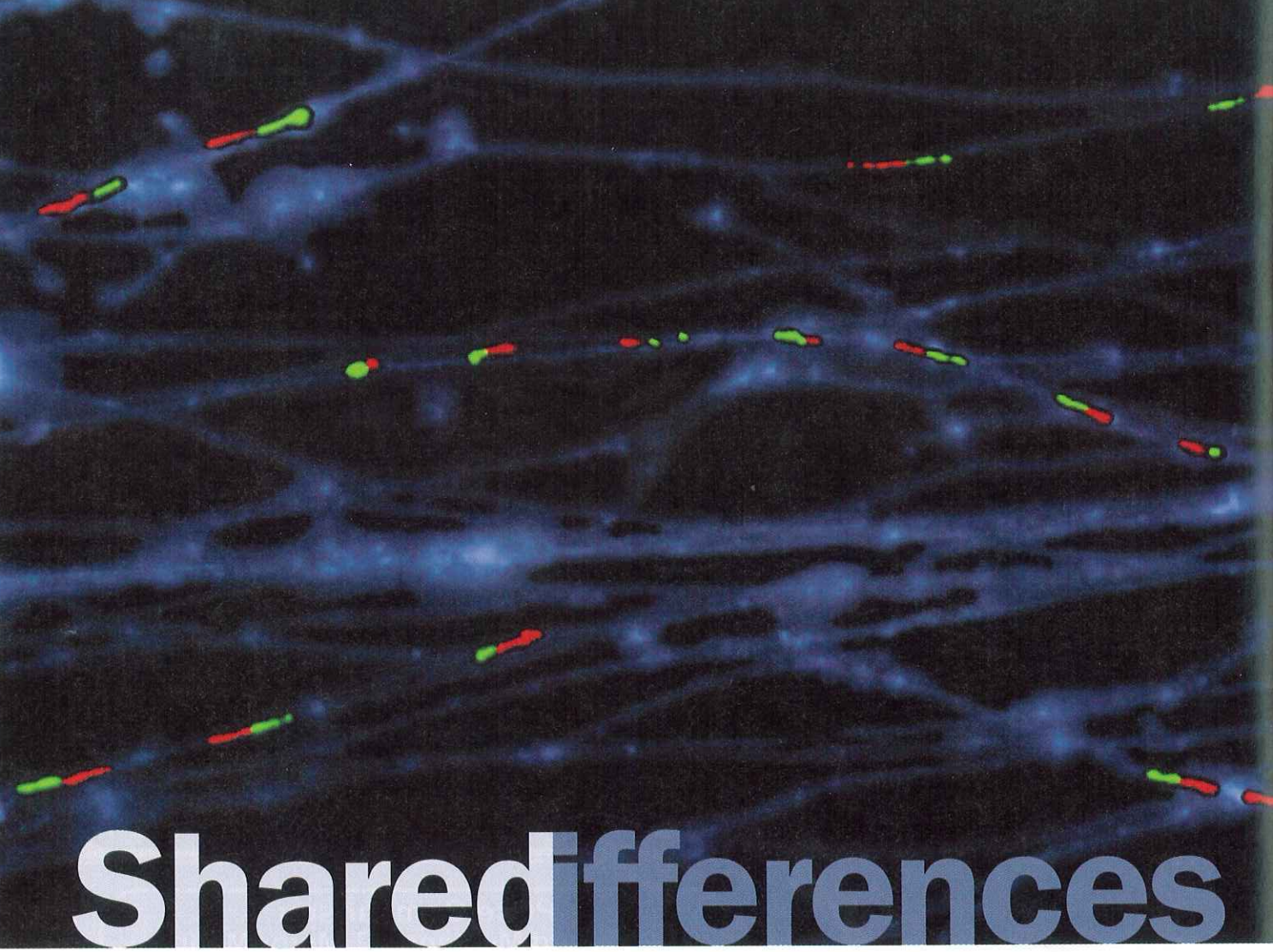


Sharedifferences

The architecture of our genomes is anything but basic

By Tina Hesman Saey

Whether you like it or not, you're a little different. If it makes you feel any better, so is everybody else. In fact, everybody is far more different than anybody had imagined.

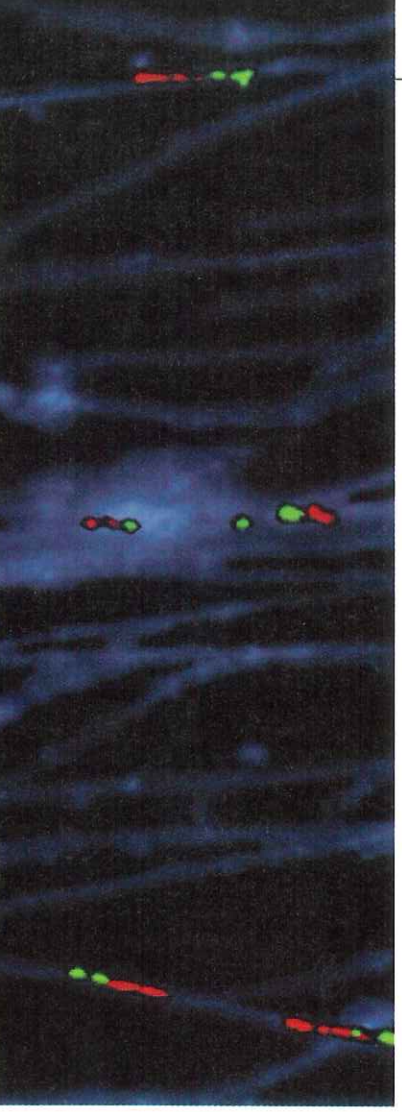
Scientists are only beginning to discover just how different humans are from each other at the genetic level and what those personal genetic attributes mean for health, history and the human evolutionary future.

It's true that people are 99.9 percent alike, if only minor spelling variations in the genetic instruction book are taken into account. In each person, about one in every 1,000 DNA bases — the chemical letters of the genetic alphabet — differs from the generic human construction and operating manual. So, on average, one person will differ from another at about 3 million of the 3 billion letters in the human genome. Researchers have recently mapped many of these single letter variations, called single nucleotide polymorphisms or SNPs, looking for variants that might play a role in complex diseases such as heart disease, diabetes and high blood pressure (*SN: 6/21/08, p. 20*).

So far, SNPs have been associated with many diseases, but SNPs can also be protective. And those little spelling differences may contain information about a person's geographic ancestry — just as whether people write *color* or *colour* is a clue about whether they hail from the United States or Great Britain.

But SNPs aren't the whole story of human-to-human distinctions. Scientists now know that a different type of variation, previously thought to be rare, is surprisingly widespread.

New research shows that the human genome has undergone extensive editing, much more sweeping than the minor letter differences or spelling variations. Entire pages may be torn from, or even stuffed into, an individual's genome. Paragraphs can be duplicated multiple times, swapped with other passages, written backward, deleted, truncated or otherwise altered in myriad ways. These differences are known collectively as structural variation, and as little as 5 percent or as much as 18 percent of the human genome may be affected.



Glowing strands of DNA from five people highlight differences among humans in the number of copies of amylase genes, which encode enzymes that break down sugars. Red and green probes bind to regions hosting the genes, and each DNA strand has a different number of the genes on the short arm of chromosome 1.

Scientists previously knew that having extra copies of an entire chromosome could lead to disorders, such as the third copy of chromosome 21 that causes Down syndrome. Research had also identified very large deletions that remove so much of a chromosome that the void can be seen under a microscope, and had revealed nips and tucks that remove single genes or parts of genes.

But until completing the Human Genome Project, an effort to map all the genes and surrounding DNA found in people, researchers had no way to detect structural changes too small to be seen under the microscope and too big to be detected by looking at individual genes.

Because these variations cover large expanses of the genome rather than single DNA bases, current estimates suggest that structural variation may encompass four times as many bases as SNPs do, meaning that people's genomes differ by an additional 0.5 percent. So any two people are really only about 99.4 percent alike.

New findings indicate that a substantial portion of otherwise healthy people are missing large chunks of their genomes, gaps that can predispose them to certain diseases. Already, structural variants, especially the type known as copy number variants, have been linked to neurological disorders such as schizophrenia and autism, to susceptibility to HIV infection, to Crohn's disease and even to tendencies in weight.

"Our work in structural variation is showing that no one is really normal," says Charles Lee, a cytogeneticist at Brigham and Women's Hospital and Harvard Medical School in Boston. Lee was among the first to discover the broad range of structural variation in the human genome.

Not necessarily two copies

Lee and his colleagues knew from work begun decades ago that some parts of the human genome contain multiple copies of certain genes. For example, light-sensitive opsin proteins, made in the eye and necessary for color vision, are encoded by a cluster of two to nine genes on the X chromosome. Some of the copies encode proteins that are better at sensing green light, while others are specialized for red. The number of copies of the genes affects how well people see colors, and missing certain ones can lead to color blindness.

In another example, each person has many copies of the immune system HLA genes. And blood disorders known as thalassemias arise when copies of genes that encode subunits of hemoglobin, blood's oxygen-carrying molecule, are missing.

Those examples were thought to be exceptions to the rule that each gene is inherited as two copies, one from the mother and the other from the father. No one suspected that parents routinely pass along three, four or more copies of entire parts of the genome, or sometimes fail to pass along a whole section.

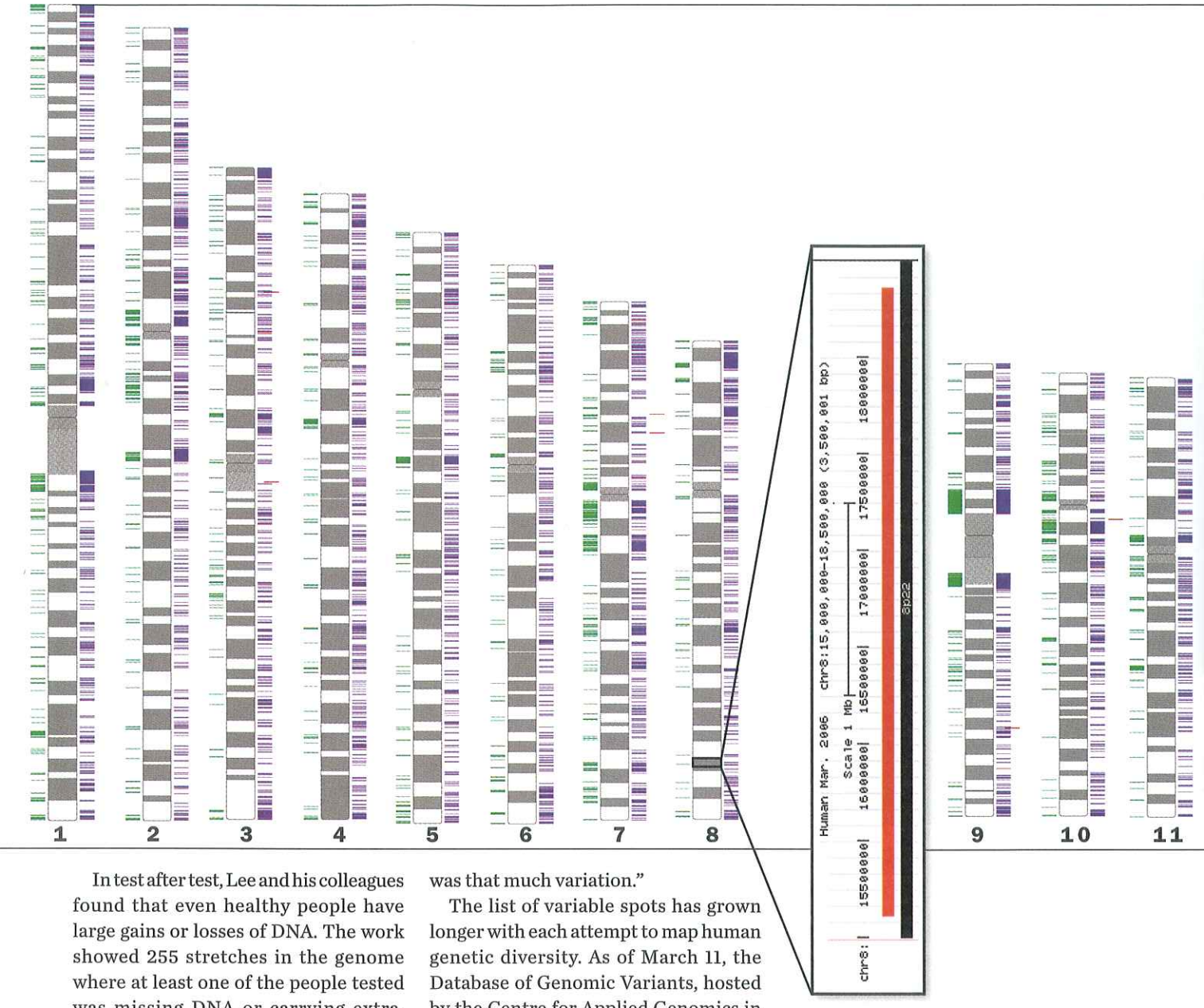
Because structural variation alters entire sections of a chromosome, the genes within those sections can be copied multiple times, inverted or deleted. As a result, the number of copies of the genes in an altered stretch can vary in many ways.

Even the Human Genome Project was affected by the assumption that most of the genome contains only a single copy of each parent's DNA, Lee contends. What the project compiled is an averaged human genome, a sequence homogenized from multiple people that represents no real person. About 66 percent is from an anonymous man of European descent. The rest is a mishmash of DNA sequences from several other people.

Left out in favor of a global template is the diversity among single letters and overall structure originally present in each of the DNA donors. Multiple copies of genes look alike or have minor differences that could be easily attributed to glitches in the decoding process. So piecing together the whole genome sequence from snippets of data ended up collapsing what should have been several gene copies into a single gene, Lee says.

He discovered structural deviation from this generic sequence while testing a method for detecting abnormalities in the genomes of tumor cells. Cancer cells are notorious for deleting, rearranging and duplicating parts of the genome. Lee wanted to map those changes by comparing the cancer genomes with the Human Genome Project consensus sequence. But first he needed to be sure that DNA samples from healthy people matched the consensus sequence.

FROM THE AMERICAN JOURNAL OF HUMAN GENETICS, JANUARY 2008



Uniquely schizophrenic
Large deletions from chromosome 8 could be linked to schizophrenia. In one patient, a portion of the chromosome was almost entirely missing: The black band shows the normal width of this region and the red band shows how much was missing from this person's chromosome.

In test after test, Lee and his colleagues found that even healthy people have large gains or losses of DNA. The work showed 255 stretches in the genome where at least one of the people tested was missing DNA or carrying extra, Lee's team reported in *Nature Genetics* in 2004. The same year another group independently found 221 places in the genome where copy number varies. Led by Michael Wigler of Cold Spring Harbor Laboratory in New York, the group reported in *Science* that, on average, any two people differ at 11 of those places.

The human genome had revealed an unexpectedly large variability.

"It was a major revelation," says Andrew Shelling, a geneticist at the University of Auckland in New Zealand. "Certainly when it was brought to our attention, we were staggered that there was that much variation."

The list of variable spots has grown longer with each attempt to map human genetic diversity. As of March 11, the Database of Genomic Variants, hosted by the Centre for Applied Genomics in Toronto, listed 6,558 locations in the human genome where variations may occur. Some locations have multiple variations – copy number can vary in many different ways within one person's genome or from one person to another.

The database lists 31 separate studies that document 21,178 copy number variants of DNA segments 1,000 bases or longer; 499 inversions (places where a stretch of DNA is spelled backward); and 16,729 insertions or deletions ranging in size from 100 bases to 1,000 bases. Many smaller variations may exist that current methods can't easily detect.

One recent effort to map structural variability and single letter changes in 2,500 people suggests that 65 to 80 percent of the population has copy number variants that run at least 100,000 bases long. About 5 to 10 percent of people carry copy number variants longer than 500,000 bases. Research-

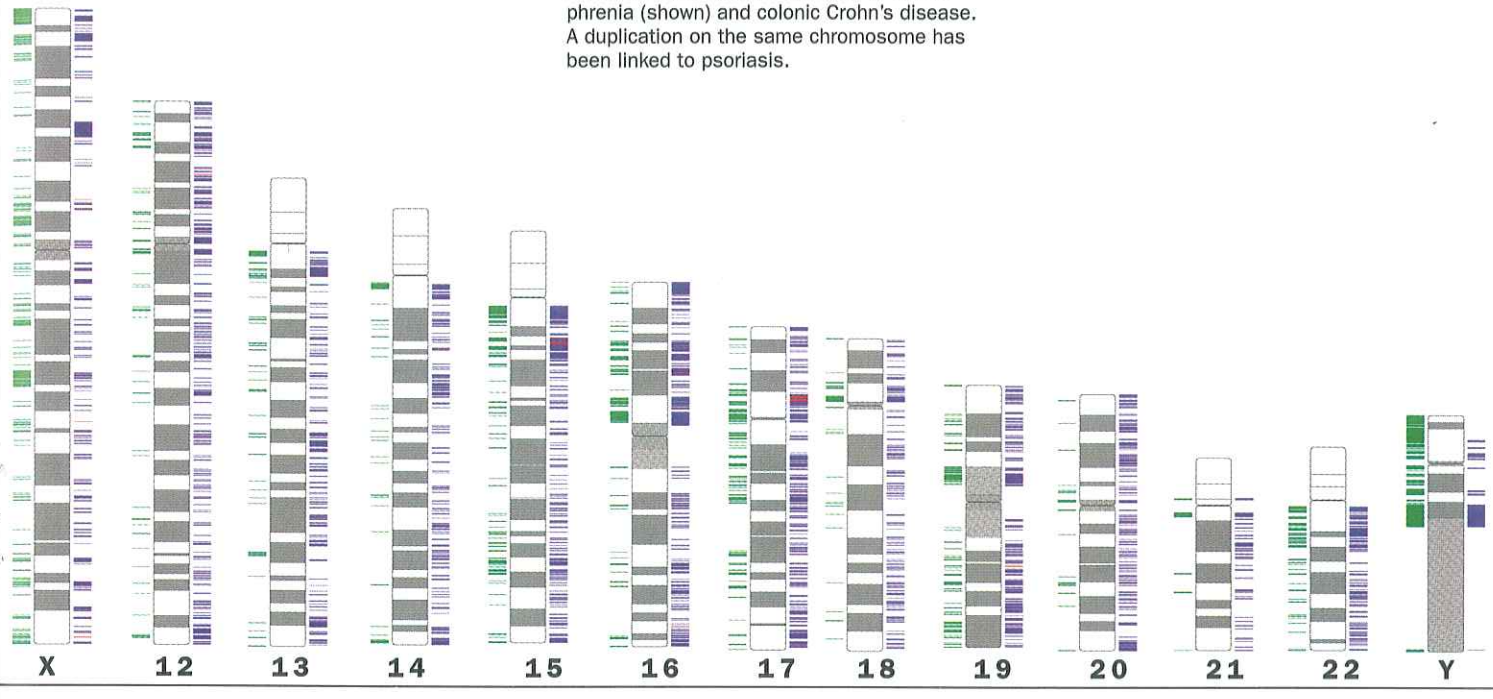
Variation in our chromosomes

A view of all the human chromosomes according to the Database of Genomic Variants reveals that people do not necessarily inherit two copies of every gene. Instead, entire sections of a chromosome (shown in blue) spanning thousands to millions of DNA bases can be duplicated, deleted or inverted, causing the number of copies of genes in those stretches to vary.

New work shows that this widespread variation within the human genome is more common than thought and is a legacy of human evolutionary history. Studies are also beginning to connect copy number variants with specific diseases. A few examples: Variation in copy number of immunity-related genes on chromosome 17 have been linked to susceptibility to HIV/AIDS. Deletions in chromosome 8 have been linked to schizophrenia (shown) and colonic Crohn's disease. A duplication on the same chromosome has been linked to psoriasis.

Key

- Reported copy number variants (Stretches where the chromosome is duplicated, deleted or inverted.)
- Duplicated segments of chromosome
- Reported end points between stretches of inverted sequences



DATABASE OF GENOMIC VARIANTS; INSET: DAVID B. GOLDSTEIN/ DUKE UNIV., MODIFIED FROM *PLOS GENETICS*, FEBRUARY 2009

ers led by Evan Eichler, a Howard Hughes Medical Institute investigator at the University of Washington in Seattle, reported in the Feb. 13 *American Journal of Human Genetics* that 1 to 2 percent of people have copy number variants, or CNVs, that are more than 1 million bases long.

"There's a real CNV burden in the population," Eichler says.

Culprits in disease

Some of these variations — and perhaps simply the volume of changes — in the genome may predispose people to diseases, other studies suggest.

A report published February 6 in the online journal *PLoS Genetics* showed that patients with schizophrenia have some of the largest variants: chromosome deletions that remove 2 million or more bases. Such megadeletions occurred only in eight of the 1,073 schizophrenia patients screened, but in none of the 2,492 healthy people in the comparison group.

Some big duplications were also found in people with schizophrenia, but healthy people had large duplications too. So researchers think that losses of large chunks of DNA, not additions, are more likely linked to schizophrenia.

The international team that conducted the study did not find evidence linking any of the common SNP variants previously implicated in schizophrenia to the disease. The results could mean that rare, harmful differences are more important in schizophrenia than are more common variations. This idea represents a change in thinking about the genetics of relatively common diseases.

Researchers have thought that common genetic variations might contribute to common diseases, such as heart disease and diabetes, when triggered by environmental factors such as poor diet, smoking and lack of exercise. This study's result indicates that, at least for schizophrenia, rare variants that remove different parts of the genome in different people might result in the same disease.

Other researchers have demonstrated that copy number variants, especially deletions, are important in autism. Last June, researchers from the University of Chicago and colleagues, writing in *Biological Psychiatry,* reported finding 51 copy number variants — including duplications as well as deletions — in 46 of the 397 people with autism in the study. Forty-one people had one variant and five people had two variants. These

variations weren't found in a healthy control group. The authors say many different variants might lead to autism, including variants that affect genes important for brain development and function, as well as variants that affect multiple other genes.

At least 14 diseases have been associated with copy number variants in the past few years, Lee says. "We're going to find a lot more out there," he predicts.

But variety is not all bad. "The vast majority of those differences will have no impact at all on the person," Shelling says. And some may even be good.

A human variety

Indeed, some variations may be tied to qualities that distinguish humans from other primates. Certainly, structural variation is part of what makes each person different from another. Varying copies of certain regions of the genome may be evolution's answer to making sure humans don't grow genetically stagnant.

"Perhaps it is a deliberate mechanism to ensure that our genome is always changing," Shelling says. "It was a bit of a shock that we could have a mechanism that would change the numbers and mess with our genomes that way."

Although humans vary at millions of bases, says Sarah Tishkoff, a geneticist at the University of Maryland in College Park, people are overall very alike.

"We still think that as a species we're still really quite similar," she says. "Any two of us are more similar than any two chimps."

Copy number variations also occur in humans' ape cousins and may be a legacy from an ancient ancestor. About 8 million to 12 million years ago, a common ancestor of humans and African great apes experienced an explosion of duplications in its genome, Eichler's group reported in the Feb. 11 *Nature* (*SN: 3/14/09, p. 14*). The large increase in the gene duplication rate continued into the common ancestor of humans and chimpanzees. Later the rate slowed again.

Those duplications may have established a genome architecture that predisposes humans to changes in copy number even today, Eichler suggests.

In the Feb. 13 *American Journal of Human Genetics*, Eichler's group describes hot spots for copy number variation, sites located in or next to duplicated regions of the genome. The locations hint that such variations arise when chromosomes line up to swap genetic information during recombination. Recombination occurs when chromosomes are paired before separation during the production of egg or sperm cells. But sometimes existing gene duplications can cause the chromosomes to misalign, and that misalignment leads to further duplications or deletions. Last year in *Nature*, Eichler and his colleagues reported comparisons of individual human genomes suggesting that misalignment and unequal swapping during recombination create nearly half of copy number variants.

> "It was a bit of a shock that we could have a mechanism that would ... mess with our genomes that way."
>
> ANDREW SHELLING
> UNIVERSITY OF AUCKLAND

Damage to DNA may lead to structural changes too. Toxins and other stressors sometimes break chromosomes. The process that repairs broken chromosomes may add or subtract DNA or mistakenly relocate a bit of one chromosome on to another.

Still other copy number variants may arise when DNA replication is stalled. And mobile bits of DNA — commonly called jumping genes — do their share of shaping the genome, too.

Most methods of detecting structural variation have not identified the precise location of the end points of the variants, information that is needed to determine the mechanism that created the variant, says Steven McCarroll, a geneticist at the Broad Institute in Cambridge, Mass., and the Harvard Medical School in Boston.

Scientists do know that most of the copy number variants in the general population are old. In a study published last October in *Nature Genetics*, McCarroll and his colleagues found that over 90 percent of the copy number differences among people in the study are due to the usual suspects, variants that are quite common. Only 8 percent of the variants identified in the study were rare, found in a single family or individual. And only 10 rare variants of the 1,320 analyzed were found in a child but not the parents, indicating a new change.

New mutations are more common in sporadic cases of disease than in cases where the patient has a family history of the disease, McCarroll says. But researchers still don't know whether rare or common variants are more likely to be associated with disease.

About 1 to 2 percent of autism and schizophrenia cases have been associated with rare structural variants, but no common variations have been found to explain inheritance of those disorders, McCarroll says.

In contrast, some cases of Crohn's disease — an inflammatory bowel disorder — have been linked to a common variant that deletes defensins, genes involved in protecting against invading bacteria. And an international consortium reported January 4 in *Nature Genetics* that body mass index is linked to a common structural variant: a deletion covering a stretch of DNA more than 45,000 bases long and containing an important gene called *NEGR1*, which is associated with obesity. No rare structural variants have been linked to either Crohn's disease or body mass, McCarroll says.

Despite what specific genetic additions and subtractions may contribute to such diseases, environmental factors may be just as important in determining who gets sick, says Shelling.

"You are not predestined by your genes," Shelling says. "If you've got a gene that protects you from having cancer, it doesn't mean you should smoke. If you've got a gene that makes you slim, you still shouldn't eat a McDonald's every night." ■

Explore more

- Charles Lee lab: www.chromosome.bwh.harvard.edu