GENOMICS

1000 Genomes Project Gives New Map Of Genetic Diversity

Talk about inflation. A decade ago, one human genome was the goal. Now nothing less than 1000 will do. By sequencing hundreds of human genomes, the 1000 Genomes Project has produced the most detailed catalog of human variation ever: a compendium of millions of previously unknown single-nucleotide polymorphisms (SNPs) and other variants. This treasure chest of genetic data, which was generated by three pilot projects, is described in the 28 October issue of *Nature*. Researchers are already using those data to pinpoint DNA involved in both complex and inherited diseases. "The resource should have a large impact on medical genetics," says Sekar Kathiresan, a cardiovascular geneticist at Massachusetts General Hospital in Boston.

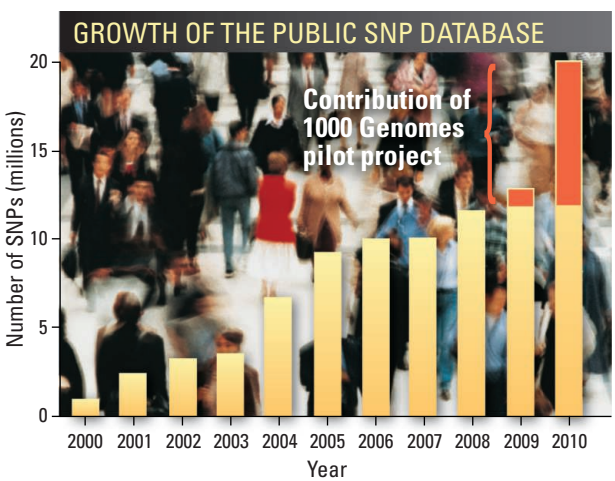
Also, on page 641 of this issue of *Science*, a second analysis describes an approach for determining another aspect of genetic variation that arises when genes and other stretches of DNA are duplicated. There is growing interest in these so-called copy number variants because of their potential ties to disease risk (*Science*, 7 September 2007, p. 1315).

Although all humans share 99% of their DNA, the relatively few differences among us matter—for disease, personality, and other traits. The first large-scale effort to identify these differences was the International HapMap Project, which identified 3.5 million SNPs, places on the genome where one base might vary from one individual to the next. Geneticists have since used those SNPs in so-called genome-wide association studies (GWAS) to home in on genes important to diabetes, heart disease, various cancers, and age-related macular degeneration, to name a few. But these GWAS covered only variants that existed in at least 10% of humans and often helped pinpoint a region but not the particular base causing the disease risk. The 1000 Genomes Project addresses these issues.

In 2008, spurred by rapid advances in sequencing technologies, the Wellcome Trust Sanger Institute in Hinxton, U.K.; the National Human Genome Research Institute in Bethesda, Maryland; and BGI in Shenzhen, China, embarked on the 1000 Genomes Project with the goal of identifying the SNPs that are present at 1% or greater frequency among humans, as well as other variants. To figure out how to do this, they set up the three pilot projects now being reported. Ultimately, the project will sequence to varying degrees 2500 samples from 27 populations around the world. The pilot data are already publicly available.

In one pilot, the researchers thoroughly sequenced the genomes of two families—mother, father, child trios—revealing that offspring inherit about 60 mutations that arose in their parents, with slightly more coming from the father. The second project involved sequencing in less detail the genomes of 179 people with ancestry from Europe, East Asia, or Africa. In the third, the consortium determined the sequences of 8140 exons—covering 906 genes—of 697 people. By testing different sequencing technologies and working out ways of preparing and analyzing samples, the nine-center group has come up with reliable methods of studying variation that are rapidly becoming the field's standard, says 1000 Genomes Project co-leader David Altshuler, a human geneticist at the Broad Institute in Cambridge, Massachusetts.

The pilot projects identified 15 million SNPs—including 8.5 million novel ones—as well as 1 million small insertions and deletions and 20,000 other structural variants. "It's a huge increase over what was [known] before," says Gilean McVean, a population geneticist at the University of Oxford in the United Kingdom, who is helping to coordinate the analysis of the 1000 Genomes Project. This total represents almost 95% of the variants that are found in at least 10% of humans.



Pinning down differences. The 1000 Genomes Project has greatly increased the number of known single-base differences that can exist among people.

Using this new resource, researchers will be more likely to find the exact SNP linked to a disease, or at least will be able to get much closer to it in their GWAS. Working with 1000 Genomes data, researchers and companies are increasing the number of SNPs that can be evaluated. In the works are microarrays that will test for the presence of 5 million or more SNPs—quite an improvement over the 1 million typically assessed with current technology.

Researchers are already using the 1000 Genomes pilot to add missing SNPs to their maps of regions associated with disease. For instance, Oxford statistician Jonathan Marchini, GlaxoSmithKline, and other collaborators took this approach in a meta-analysis of 20 studies that had looked for genes linked in various ways to smoking. Those studies had implicated a cluster of genes on chromosome 15 that code for proteins that bind strongly to nicotine. As they reported in the May issue of *Nature Genetics*, the researchers used 1000 Genomes data to home in on a particular SNP that affects transcription of one of these genes. "The 1000 Genomes data set will allow many other groups to carry out fine-mapping experiments in the same way … and will help people focus in on the underlying causal variants," says Marchini.

The data can also help track down genes involved in rare genetic diseases. Kathiresan and his colleagues wanted to find the faulty gene in a family with very low cholesterol levels, in the hope that the gene could provide clues about new cholesterol-lowering strategies. "Our analysis relied on an assumption: the causal variant in this family is private to this family and therefore, the causal variant would not be present in any [existing] public databases of variation," Kathiresan explains. With the help of the new data, they narrowed their search from thousands to 481 SNPs, eventually tagging two in a gene called *ANGPTL3*, they reported 13 October in the online edition of *The New England Journal of Medicine*.

But even the 1000 Genomes Project has limitations, as its searches for genetic variation tend to skip over variation in large stretches of highly duplicated DNA. About 1000 genes lie in regions of the genome that have been duplicated. That realization prompted Evan Eichler of the University of Washington, Seattle, and

CREDITS: (SOURCE) MARK A. DEPRISTO/BROAD INSTITUTE OF MIT AND HARVARD; (PHOTO) GETTY IMAGES

his colleagues to come up with a way to analyze this previously impenetrable DNA.

As described on page 641, Eichler and his colleagues have developed a technique for counting the number of copies of a gene in any duplicated region. The genes can vary in copy number among people. The number can affect how much of that gene's protein is produced, and consequently, the function of that protein.

Eichler's team has also come up with a way of distinguishing near-identical copies. Over time, copies tend to develop slight sequence differences that could also affect how that gene—or its protein product—works. Eichler's team has cataloged these telltale variant bases for about 70% of the duplicated genes. "It's opened up a whole new area of genetic diversity that we have not been able to tap previously," says Eichler.

Surprisingly, analysis of the 1000 Genomes data showed quite large differences in the copy number of certain genes between the African, European, and Asian populations, Eichler reports. "Humans are more different than we would have ever thought," says Eichler.

"Once all this variation is revealed, it changes the way you can think about [doing] genetics," notes 1000 Genome Project co-leader Richard Durbin of the Sanger Institute. For most of the history of genetics, researchers have been fishing out variation without knowing what was there. Now, "we are right on the cusp where we do genetics in the light and [see] exactly what it is that we are studying."

–ELIZABETH PENNISI

IRAQ WAR

Leaked Documents Provide Bonanza for Researchers

The Pentagon is fuming after last week's release of a huge cache of classified Iraq War data by the organization WikiLeaks. But researchers struggling to build an accurate picture of the death toll in post-invasion Iraq are thrilled. "It is hard to overstate the significance of this development for the conflict field," says Michael Spagat, an economist at Royal Holloway, University of London, U.K.

Within the nearly 400,000 leaked documents is a stream of raw data called SIGACTS—for Significant Activities—that chronicles the casualties directly observed by U.S. soldiers in Iraq. In late summer, WikiLeaks passed a copy of these data to Iraq Body Count (IBC), a London-based organization that has tallied the war's death toll using media reports of casualties. Their numbers do not include insurgents or soldiers. And because not every violent death is reported in the media, IBC's numbers are known to be an underestimate of the true number of war dead. But how much higher the true number is has been a source of intense debate, with surveys of Iraqi households yielding a wide spread of casualty estimates (*Science*, 20 October 2006, p. 396). As *Science* went to press, the IBC toll for Iraqi civilians stood at 98,585 to 107,594 violent deaths.

According to the IBC analysis of the leaked SIGACTS data, published online on 25 October, more than 109,000 violent deaths in Iraq were logged by the U.S. military between January 2004 and December 2009. Of these, over 79,000 were civilian deaths comparable to those logged by IBC, which recorded about 91,000 over the same period. By extrapolating from a sample of the data, IBC estimates that at least 27,000 civilian deaths went unrecorded by the U.S. military, while the military observed 15,000 comparable deaths that the media missed. Most of these unreported deaths were from small incidents of violence, with between one and three casualties. This confirms a widely assumed bias in media reporting in favor of larger incidents, such as suicide bombings. "But with such a huge overlap, it does not seem very likely that there are a large number of civilian deaths missed by both sources," says Spagat, who helped IBC with its analysis.

Buried data. WikiLeaks' Julian Assange at a briefing on the release of classified Iraq war records.

CREDIT: JOHN STILLWELL/PA PHOTOS/LANDOV

Taking the WikiLeaks data into account, IBC now estimates that at least 150,000 have died violently during the war, 80% of them civilians. That falls within the range produced by an Iraq household survey conducted by the World Health Organization—and further erodes the credibility of a 2006 study published in *The Lancet* that estimated over 600,000 violent deaths for the first 3 years of the war (*Science*, 18 January 2008, p. 273).

The leaked data are sure to keep researchers busy for months to come. Besides the number of casualties, the SIGACTS release includes geographic locations of the violence and other information that has not been available until now. But there could be serious challenges to those hoping to publish an analysis, says Gary King, director of the Institute for Quantitative Social Science at Harvard University. "I have had a couple of students asking [Harvard] for permission to use the previous WikiLeaks data release, and last I heard they still weren't allowed to touch it." But others are more optimistic. "As long as the data is stripped of information that could be used to identify anyone, it shouldn't be a problem," says Christian Davenport, a political scientist at the University of Notre Dame in Indiana who studies conflict mortality. "What's important is that people appreciate the complexity." For example, he says, "we don't know exactly how these data were gathered." It represents the results of an experiment, "but we don't know the methods."

–JOHN BOHANNON