Drupal.behaviors.print = function(context) {window.print();window.close();}>

All the Variation

October 03, 2011

All the Variation

By Ciara Curtin

People are very different from one another and yet they are so similar. The concept of human diversity is relative — diverse as compared to what? A good comparison may be to chimpanzees. According to Sarah Tishkoff, an associate professor at the University of Pennsylvania School of Medicine, chimpanzees are roughly three times as diverse as humans. "People are often surprised by that because there's a lot more modern humans around than there are chimps. But it means that in the past, chimps had a larger population size than humans," she says. "It means that at some point we had a major restriction in population size — or what's referred to as a bottleneck — during human evolution and we lost a lot of the diversity during the speciation process."

Currently, modern humans are thought to have originated in Africa about 200,000 years ago. Then, around 50,000 to 100,000 years ago, some began to leave Africa. Whether this migration occurred once or in multiple waves is unclear.

"We're a relatively young species and the divergence of human populations around the world is relatively recent," says Noah Rosenberg, a population geneticist at Stanford University. "The magnitude of the genetic differences among human populations in different parts of the world is smaller than the magnitude of the genetic difference among different populations of many organisms."

Again, as compared to chimpanzees, people are similar to each other at the genetic level, though not all of the differences that may be lurking in the human genome are known. "We differ at about one out of every 500 to 1,000 nucleotides, if you were to just look at single nucleotide polymorphisms in the genome," Tishkoff says. "If we were to consider structural variation — like inversions, insertions, and deletions — we're just beginning to understand how much we differ."

It is clear, however, that there is variation that makes people who they are. "As we start collecting data from next-gen sequencing, ultimately you get all the rare variants and clearly there is variation out there that is making each of us unique," Tishkoff adds. "It's not that there isn't variation out there, but you will find there is some that's going to be specific first to the individual, [and] there may be some that [is] specific to populations or regions."

Understanding the diversity that exists among people can help researchers retrace the evolution of the species, understand its population structure, divine the role of rare variants, and get a firmer grasp of medical genetics.

Genome-wide association studies and large sequencing efforts are helping fuel that basic and medical research. Last year, the 1,000 Genomes Project finished its pilot phase, for which the group sequenced around 180 people at low coverage as well as the exomes of about 700 people. From this, the group found 15 million SNPs, 1 million indels, and 20,000 structural variants, which it said accounts for 95 percent of human variation. Other projects have loftier goals: the Personal Genome Project aims to sequence 100,000 people. Soon, there will be an overload of large data sets.

"It's a huge opportunity. ... Population genetics has been waiting for these kinds of data sets for decades," says Graham Coop, an assistant professor at the University of California, Davis. "It's a huge opportunity to learn more about population history, particularly in humans."

Another level of scale

Researchers are beginning to use sequencing and other large-scale genomic analyses to understand human variation and human evolution. In her lab, Tishkoff studies normal variable traits, often in indigenous populations. One such trait is lactase persistence. In 2007, she and her team reported in *Nature Genetics* that lactase persistence in Africans and Europeans is due to convergent evolution. "For that we focused on a region near the lactase gene and we did look at SNPs going out several million base pairs, but at the time, the technology wasn't really developed to do the more wide-scale genome-wide approaches," Tishkoff says. She and her colleagues found three SNPs in Tanzanian, Kenyan, and Sudanese people associated with lactase persistence that differed from the SNP for the same trait found in Europeans.
[pagebreak]
For future studies, "we're interested in using genome-wide SNP data, next-generation sequencing data, transcriptome data, and so on," she says. "We're using an integrative genomics approach to study population history, population relationships, and genetic basis of adaptation to different environments in Africa."

"It's not just us," she adds. "Everybody's at this very beginning stage of trying to correlate variation on a genome scale with phenotypic variation."

More recently, her group took a targeted genetic approach to study nucleotide variations linked to malaria susceptibility. As they reported in the *American Journal of Human Genetics*, the researchers focused on the nucleotide diversity of the glycophorin gene family in a number of African populations. They found that these gene family members had high levels of nucleotide diversity and gene conversions. In addition, they identified a haplotype that leads to three amino acid changes in the extracellular domain of glycophorin B, which appears to have evolved in populations with high exposure to malaria.

"We had to use allele-specific amplification to be able to target each of them. Standard NGS methods are going to get these short little reads that's going to be extremely difficult to properly align duplicated gene regions," she says. For other studies of variation in populations exposed to malaria — and other environmental factors — Tishkoff and her

team are trying to correlate the variation they see to those outside influences. "That might give us a clue about either genes that play a role in adaptation, or we may identify particular variants that are important and they may not be otherwise easily detectable," she says. Infectious diseases, like malaria, have exerted strong selective forces on humans, she adds, but so have diet and environment.

Tishkoff is not alone in beginning to bring SNP and other data together to study humans. Rosenberg's group at Stanford also studies human evolutionary history using such genetic information. In a 2008 *Nature* paper, Rosenberg and his colleagues report surveying genotype, phenotype, and haplotype variation in worldwide human populations. He and his team analyzed nearly 526,000 SNPs and roughly 400 CNVs in 29 populations. "We were interested in comparing the inferences about human population relationships from different types of genetic markers and what we found in that study was that CNVs, which had not previously been studied in diverse worldwide populations, could actually be used in order to understand population relationships," Rosenberg says. "What we found was that CNVs do contain information about human evolution and human migrations."

Rosenberg's group also found that individual haplotypes tended to be specific to certain regions. "They tend to be less widely distributed and they tend to be more diagnostic of particular regions of origin, and so one of the results of our study was that we found that use of haplotype information can lead to advances in the study of human evolution," he says.

This, Rosenberg adds, is of particular interest these days as researchers debate the role of rare variants in determining risk of complex disease. "Rare variants tend to be less widely distributed than common variants and more distinctive to particular populations, so the more interest the community has in rare variants, the more important it's going to be to study populations in different locations that will have distinct combinations of rare variants," he says. "If rare variants are going be really important, then we need to look at them in a lot of different populations."

Across Europe

Other researchers are trying to map how gene variants have spread geographically. To make sense of the data from GWAS and other genomic studies, some population geneticists have turned to principal component analysis. That approach can be used to summarize large data sets, Rosenberg says, and has been used since the earliest multi-locus studies were done in the 1970s.
[pagebreak]
"Principal component analysis is a tool for exploring multi-locus population genetic data, and representing large-scale population genomic variation in a small number of dimensions," adds Olivier François, a professor of applied probability and statistics at the Grenoble Institute of Technology in France. "The development of GWAS has made PCA very popular for correcting tests for spurious population structure. One reason of the popularity of the PCA method is that it is simple and fast."

Indeed, researchers have applied this approach to study everything from the rise of pastoralism to the origin of *Helicobacter pylori* in the human gut. It can also be used to create synthetic maps, like ones that correlate variant gradients to geographic expansion of populations. "Standard explanations have tried to correlate gradients in PC maps to origins of geographic expansion," François says. However, he adds, some recent work

has shown that those gradients can be seen without making any assumptions about expansion.

François and his team set out to test whether the gradients could be detected in simulated data under various expansion scenarios, as they report in Molecular Biology and Evolution. For example, one axis of variation in Europe, called PC1, was thought to support the idea that farmers expanded through Europe to replace hunter-gathers with little interbreeding. Instead François' group found that the variant gradients are perpendicular to the direction of expansion. "Let us suppose the farmers expanded in Europe from the southeast during the Neolithic and replaced the resident populations of hunter-gatherers. If this expansion left a footprint on European genetic variation, then this signature should be visible in PC1, with gradients oriented in the Southwest-Northeast direction," François says. "We thus expect the largest differentiation between populations from the southwest and the northeast."

"Our work also shows that a small amount of genetic admixture with resident populations — incomplete replacement — could change the orientation of PC maps drastically," he adds. "So, our work points out severe difficulties with interpretation of PC maps."

More recently, his team used a separate approach to reproduce the geographical spread of a haplotype. Focusing on the Y chromosome haplogroup r1b1b2, which is common throughout Western Europe, the researchers used a wave-of-advance model, a standard model for the advance of farming in Europe, and found that the expansion was less than 10,000 years old — and could be as recent as 3,000 years old. "We expect that ancient DNA will give us a clearer answer to this question in the near future," he adds.

Randomness

But some of the differences seen between human populations are likely just due to chance. "We tend to think about differences as allele frequencies, the frequency of an allele at a single nucleotide polymorphism. And the vast majority of allele frequency changes are probably due to genetic drift, and so they are not changes in response to the external environment, but just to the randomness of reproduction," Davis' Coop says.

Indeed, he adds, allele frequencies within and outside of Africa differ — mainly because of the separation times of those populations and limited migration between them.

Coop's group has taken advantage of newly available data sets to study human populations. Using the Human Genome Diversity Panel genotyping chip, he and his team searched for SNPs with large differences in allele frequencies between populations, as they report in *PLoS Genetics*. "What we were interested in was the SNPs that showed particularly large allele frequency differences between populations, as those are a priori more likely to be as a result of selection," he says.
[pagebreak]
What they found, though, was that only a small number of loci vary substantially between different populations. "Selection has actually been relatively inefficient at creating differences between populations," Coop says. "There's only a relatively small number of loci that strongly differentiate populations, which seems to indicate that selection hasn't been rampant in the human genome."

But he adds selection has clearly occurred — humans have adapted to environments the world over. "There have been genetic changes between populations, but that's not linked to single loci," he says.

This work, Coop notes, was based on a genotyping chip, so it does not capture all of the variation in the human genome. "Things like the 1,000 Genomes Project are going to be a huge opportunist to study an unbiased ascertainment of variation within a population," he adds.

Regulation

Then, of course, there is another level at which human populations can differ: how they regulate genes. "[I am interested in] trying to understand the function of the genome," says Stephen Montgomery from Stanford's School of Medicine. "If you can relate these discovered variants of different classes to an effect like gene expression, you can understand, potentially, how the genome is influencing a cell and defining more complex traits." Then by looking in different populations, Montgomery can try to predict whether what he finds in one population will be relevant to another, based gene expression.

In a *PLoS Genetics* paper published this summer, Montgomery — then at the University of Geneva Medical School in Switzerland — and his colleagues surveyed the landscape of regulatory variation using data from the 1,000 Genomes Project and HapMap3 cell lines. In particular, they focused on non-synonymous variants to see whether they were differentially expressed.

"On different haplotypes we see different levels of expression for those alleles and we can see, actually, that a considerable amount of nonsynonymous SNPs were differentially expressed," he says. "If we can actually integrate the expression information with identification of pathogenic mutations, you get a better understanding of any individual's disease-risk profile or prevalence of a particular type of trait." If someone has a deleterious variant that's not being expressed, then it's likely not having a large impact on that person, he adds.

Then last month in the *American Journal of Human Genetics*, Montgomery's team reported on interactions between regulatory and coding variation. Again drawing from 1,000 Genomes data, the team reported that genomic variation is shaped by cis-regulatory variation — that is, regulatory and coding variation modify each other.

The data

While the availability of new and larger data sets is a blessing, it also can be a curse. Choosing which data to focus on is a difficult choice. "There is so much data being generated that it is difficult to choose problems to work on that are going to be timely. The rate at which data is being generated, data sets are quickly dwarfed in size so that a particular data set may seem big, but then six months later, there is something much bigger," UC Davis' Coop says. "It's easy to get swamped by the sheer size of data sets being produced."

Similarly, Stanford's Rosenberg says developing tools to work with such quantities of data can be a tough job. "With large genomic data sets we have information on sequences or haplotypes or genome-wide markers these data potentially contain a lot more information

than classical data sets that only had a small number of markers, and it's a challenge to develop tools that can take advantage of the new features of the data," he says.
[pagebreak]
To integrate various data types, as many researchers hope to do, is also a statistical problem. Members of Montgomery's lab at Stanford hope to be able to make predictions about individual risk, based on expression profiles and other information. "It's a lot of information that has to be integrated and it really is a statistical challenge in some regards, and also a challenge based on the prior understanding of disease and its mechanism," Montgomery says.

And to trace back human lineages to see when populations diverged, researchers also need better statistical tools and newer computer simulations. For example, UPenn's Tishkoff says, nearly every population in Africa is admixed, and that can affect estimates of population divergence. "[It] complicates the analyses when you try to infer when two populations split. Say that each of them admixed with a third population — that's going to throw off your estimates," she says. "There's got to be a development of better statistical tools, better computational approaches — that's going to be critical as well."

SIDEBAR

Missing out

Noah Rosenberg and his colleagues reported in *Nature Reviews Genetics* last year that about 75 percent of GWAS studies exclusively studied European populations. This year, in a *Nature* report, Carlos Bustamante, Esteban González Burchard, and Francisco De La Vega put that number at closer to 96 percent. By excluding diverse populations from studies, both science and non-European populations may be missing out.

From a scientific perspective, not including populations other than Europeans means that the full scope of human variation is not known. "We know that variation is somewhat different across different geographic populations, different geographic regions and it is important that if we are going to get a full view of human genetic diversity and understand how that affects phenotypes and disease that we actually go out and look other populations other than Europeans," says Graham Coop at UC Davis.

UPenn's Sarah Tishkoff notes that since there is so much diversity within Africa itself, populations there may have variants that cannot be found anywhere else.

"Some of the factors that motivate the study of genetic risk factors worldwide include the possibilities that risk variants will differ in frequency around the world, risk variants may differ in effect sizes, and different risk variants may be present in different populations. And each of these factors has known examples, so it's desirable to have genomic studies performed in as diverse a collection of populations as possible," Stanford's Rosenberg adds.

Then, if those other risk variant and risk factors are not known for a variety of populations, those populations may not benefit from genomics research and the personalized medicine it may bring. "We would like to make sure that the results of genomics studies are valid for the worldwide population as a whole," Rosenberg says.

Related Stories

- Personalized Perspectives
  May 1, 2011 / Genome Technology
- To Stop Cancer
  April 1, 2011 / Genome Technology
- Fifth Annual Young Investigators
  November 30, 2010 / Genome Technology
- Over the Hurdles
  September 1, 2010 / Genome Technology
- Systems Biology Fights Cancer
  March 31, 2010 / Genome Technology