MOLECULAR ANTHROPOLOGY

How to Sample the World's Genetic Diversity

Allan Wilson wasn't at the meeting at Stanford last month—he died last July—but, without invoking the supernatural, it's safe to say his presence was strongly felt. The meeting was the first in a series to plot the course of an ambitious project that Wilson, a molecular anthropologist at the University of California, Berkeley, and Luca Cavalli-Sforza, a population geneticist at Stanford, conceived last year: to survey the genetic diversity of humanity. By analyzing how people vary one from another, researchers hope to glean clues into human origins, evolution, migration patterns, adaptation, and disease. The goal, says Berkeley geneticist Mary-Claire King, one of the organizers of the meeting, is nothing short of understanding "who we are as a species and how we came to be."

There was little time to waste if those questions were to be answered, Wilson and Cavalli-Sforza realized, as many of the aboriginal tribes most essential for reconstructing human history are disappearing—some literally dying out, but most lost to acculturation, intermarriage, and the influx of modern society. Wilson, Cavalli-Sforza, and colleagues issued a plea for help last year in the journal *Genomics*, calling on scientists worldwide to help them collect DNA samples from hundreds of indigenous populations and preserve them for future study, creating a huge resource and database for the scientific community (*Science*, 21 June 1991, p. 1615).

But there was a problem: Though Wilson and Cavalli-Sforza were united on the project's goal, they were deeply divided on how to achieve it. Cavalli-Sforza wanted to use the traditional approach of sampling well-defined populations, while Wilson, eschewing all the assumptions inherent in identifying populations in the first place, wanted to sample along a geographic grid, collecting DNA from aboriginal peoples at more or less evenly spaced locations around the world. They reached no agreement before Wilson's death, though that has not dampened enthusiasm for the project, which has garnered an outpouring of support from anthropologists, geneticists, and linguists worldwide and, as of this summer, some modest funds for three planning workshops.

At the first of these workshops, held at Stanford in July, 40 top population geneticists and molecular anthropologists found themselves dealing with the same issue that had divided Cavalli-Sforza and Wilson, albeit framed more broadly—namely, how best to sample the world, and do so without bias. The problem, as the workshop participants were well aware, is that the way in which they collect data today will determine what answers scientists can tease from them later. And everyone wants something different. Anthropologists are seeking clues about the migration of early humans out of Africa and the settlement of the Americas. Linguists want to look at how languages and cultures move with people, and population geneticists want to evaluate the relative importance of drift versus natural selection, among other things, in human evolution.

Thus the challenge to the group, as stated by Pennsylvania State University anthropologist Kenneth Weiss, another of the meeting organizers: "Can you help us design a sample to answer different questions, and to be sure that what we propose now will be useful later?" That set off 3 days of intense debate not just on population versus geographic sampling but also on sample size, the relative merits of blood samples versus cell lines, even the goal of the project: whether it is to survey human biodiversity or reconstruct human history. (The answer, it seems, is both.)

> **"Can you help us design a sample to answer different questions, and to be sure that what we propose now will be useful later?"**
>
> **–Kenneth Weiss**

Populations versus geography. Cavalli-Sforza gamely took a first crack at outlining a strategy, lobbying hard for the population-based approach he has followed for years and dismissing Wilson's geographic grid as "impractical." The chief difference between the two approaches, he says, is that Wilson's looks at individuals in specific locations; whereas his looks at populations, defined by some ethnic identifier like language or culture, in regions. Both are useful for reconstructing human evolution and migration, and each has its strengths, said Cavalli-Sforza. A grid approach, for instance, is particularly useful for tracking the spatial advance of new mutations. But the problem with a grid on a worldwide scale, he said, is that there simply aren't representatives of aboriginal populations scattered every 50 or 100 miles across the globe. Some places may have no inhabitants at all, while others may have several tribes. "We have to consider the ethnic origin of people. It is not enough to go by a lattice," or grid-like approach, asserted Cavalli-Sforza, adding, "Unfortunately, Allan is not here to defend his view." "He'll try," quipped King, alluding both to Wilson's doggedness and the fact that many of Wilson's "descendants"—including herself—were in attendance.

Instead of using a grid, Cavalli-Sforza wants to select a minimum of 200 isolated aboriginal tribes that best represent ancestral populations. A scientific/medical team would then collect blood samples from 50 individuals in each and rush them off to a few centralized labs, where white blood cells from each person would be transformed into permanent cell lines. While cell lines are a financial and logistical nightmare, he conceded—at the worst case, they might cost $500 each, and blood samples have to be at a lab within 48 hours—they are the only way to ensure an inexhaustible supply of DNA from members of vanishing populations.

Cavalli-Sforza tried to accommodate Wilson's preferences as well, proposing that each team would also collect hair roots or cell scrapings from the inner cheek—which provide a simple and cheap, though limited, supply of DNA—or perhaps blood samples not for immortalization from additional people in the surrounding region. Even though it wouldn't establish a full-fledged grid, this strategy would allow some of the wider sampling Wilson had advocated.

But Cavalli-Sforza instantly found his compromise challenged, not just by the Wilson camp but also from others, like Weiss, who were disturbed that the number of populations was so small—a mere 200 out of the 5000 to 7000 believed to exist. "It greatly restricted the kinds of questions we could look at," said Weiss. (In all fairness, Cavalli-Sforza saw 200 as the "absolute minimum," chosen with an eye on the budget, tentatively estimated at about $23 million for the first 5 years of the project.)

The arguments did not begin in earnest, however, until Cavalli-Sforza took a stab at spelling out his criteria for selecting populations. The groups should be aboriginal, he said, which in the New World means those that were in place as of 1492, and well-defined, either by language, geography, or endogamy (the habit of marrying within one's own tribe). In addition, there should have been little recent interbreeding with other groups. Practical considerations enter in as well: A team should be able to reach the group "without spending $1 billion," said Cavalli-Sforza, which would rule out certain populations in Borneo that take 2 to 3

weeks to reach. And to ensure access, an anthropologist must already be working with the tribe.

"I am very troubled by the list [of criteria]" said Mark Stoneking of Penn State, who, like his mentor Wilson, believes the population approach is predicated on too many assumptions. "It just focuses on well-defined ethnic and linguistic groups. And when you are done with your survey you will find the human species is made up of well-defined ethnic and linguistic groups. By sampling that way you bias the results."

The crux of the issue, said Svante Paabo of the University of Munich, formerly of the Wilson lab, is how to define a population. Cavalli-Sforza believes, as do numerous other population geneticists and linguists, that language is a useful if imperfect approximation for identifying a population. True enough, said Paabo, but he worries about those instances when genetic and linguistic boundaries do not coincide. If you sample only on the basis of linguistically defined populations, he cautioned, "you may miss something interesting." An alternative would be the type of experiment he and former Wilson lab colleague Anna Di Rienzo, now at the University of California, San Francisco, are contemplating for the Nile Valley. They plan to collect DNA samples every 50 kilometers to see whether genetic variation does in fact correlate with linguistic variation, among other things.

L.L. CAVALLI-SFORZA/STANFORD UNIVERSITY

Guides to the past. The genes of indigenous people, like these Andaman Islands aborigines, can help us understand the evolutionary history of the human race.

Others grumbled about Cavalli-Sforza's admittedly symbolic focus on 1492 as the time when aboriginal groups began to be displaced. Plenty of "jostling went on before then that could have had an effect on gene flow," said Weiss. "Agriculture, which arose 10,000 years ago, had a big effect. The Romans expanded long before 1492, and those people expanded at the expense of someone."

But it was Stoneking who laid out the first major challenge when he questioned the very underpinning of the survey. We don't need immortalized cell lines, he blithely told the group. Not only are they too expensive, but the logistics preclude you from getting samples from places far from airports. Instead, he said, they could get nearly as much information by extracting DNA directly from blood samples. With existing technology, one sample would supply enough DNA for 1000 PCR procedures (a very sensitive method of analyzing DNA variation)—"basically an infinite amount." What's more, he said, whole blood is stable one week on ice, and the cost is $5 to $10 a sample, instead of $500 or so per cell line. The upshot, he said, is that they could survey 500,000 to 1 million people for the same price as Cavalli-Sforza's 10,000. That would enable scientists to recreate later whatever sample they like by selecting data from the database—for example, reconstructing a geographic grid to look at Bantu expansion, or focusing on an entire village to look at evolution on a microscale.

Stoneking's idea of "saturation" sampling caught people's fancy, but no one was willing to budge on cell lines. Says Weiss: "We are conservative enough as scientists that no one wants to give up something that we know works." And so the group remained at an impasse, chafing at the limits of such a small sample but unwilling to abandon Cavalli-Sforza's focus on well-defined groups and cell lines.

Finally, the theoreticians offered a way out when they challenged another tenet of the population-based approach: that DNA from 50 individuals per group is needed to tackle most questions about human diversity. "There has been one big sleight of hand here. Where did 50 come from? Why not 25 or 100?" asked Charles Langley, a *Drosophila* geneticist at the University of California, Davis, who spearheaded the charge.

The answer, said Cavalli-Sforza, is pragmatic: "One person can bleed 50 people and get to the airport in 1 day." That's not the only reason, adds Robert Sokal of the State University of New York at Stony Brook, who notes that a sample size of 50 is needed to give an accurate estimation of gene frequencies in a population.

Times are changing, responded Langley, who was supported by Montgomery Slatkin of Berkeley, Joe Felsenstein of the University of Washington, and others. Until recently, Langley explained, few markers existed for studying human genetic variation, so geneticists had to compensate with a large sample size. But now, he said, with an essentially unlimited supply of new DNA markers—coupled with sophisticated new statistical tools—most questions can be answered with a smaller sample by simply increasing the number of the gene loci examined, provided the genes are independent. Explained Felsenstein: "If you double the number of loci, you can halve the number of people." For most questions, in fact, a sample size of 25 or even 10 per population would be sufficient, this group argued.

Compromise reached. But there are tradeoffs. While the approach is extremely powerful for addressing global questions of evolutionary history, such as migration patterns, it is not good for questions that involve specific loci, such as how disease susceptibility is distributed among populations. "Anthropologists will have to face reality—some questions of interest just can't be answered," said Weiss, who was clearly taken with the idea of spending less on cell lines.

After some haggling, that argument carried the day, and the group settled on collecting samples from 25 individuals in each population for immortalization rather than the original 50. And that meant they could survey 400 populations instead of 200, to the delight of everyone there. The meeting was breaking up when Felsenstein, one of the advocates for small samples, got cold feet. "We think we are correct," he later explained to *Science*, "but it is a huge project. What if they go out and use our number and we are wrong?" As a "fall back," he urged them to collect many extra blood samples, as Stoneking recommended.

"I always intended to," answered Cavalli-Sforza, who made good on his word. In the meeting summary, the organizing committee agreed that in addition to establishing cell lines from a core group of well-defined populations, the goal should be to collect blood samples from many individuals in each region where the teams are working. And with that, the group crafted a compromise that even Wilson, who had vowed to dig in his heels, might have accepted. At least his former postdocs, like Di Rienzo, did. "I am happy to live with that," she said. "It provides a control to see if the definition of a population is real or in your head."

At the next workshop, however, consensus may prove more elusive. Several dozen leading anthropologists will meet at Penn State in October with their lists of populations that must be surveyed—a list that seems certain to exceed the allotted 400 populations. Setting priorities may make the sampling issues seem like a piece of cake.

–Leslie Roberts