GENOME RESEARCH

A Closer Look at SNPs Suggests Difficulties

Using the wildly popular genome markers called SNPs to track genes may be less straightforward than researchers expected

SKOKLOSTER, SWEDEN—During the past year, single-nucleotide polymorphisms, commonly referred to as SNPs (pronounced snips), have taken the genomics community by storm. SNPs are single-base variations in the genetic code that occur about once every 1000 bases along the 3-billion-base human genome. Many researchers think that knowing the locations of these closely spaced DNA landmarks will ease both the sequencing of the human genome and the discovery of genes involved in such major human diseases as asthma, diabetes, atherosclerosis, schizophrenia, and cancer. But earlier this month at the first international meeting devoted to SNPs,* enthusiasts heard sobering news.

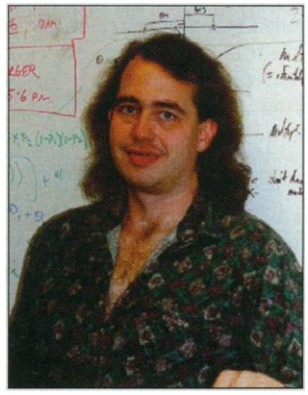
Although no one doubts that SNPs will ultimately prove to have some value in tracking disease genes and understanding human genetic diversity, new results presented at the meeting suggest that the task could prove more difficult than many had initially thought. In some cases, SNPs might fail to pick up disease genes, or researchers will need to have many more SNPs located in and around the suspected disease gene to make their case than first anticipated. Other work suggests researchers will also need more information about the history of the people being studied, such as their migration patterns, to make sense of their SNP data.

By the end of the 3-day conference, even the organizers went home shaking their heads. "There are so many problems and unanswered questions," complained Anthony Brookes, a co-organizer and geneticist from the University of Uppsala in Sweden. "At the moment, we're finding our way in the dark."

Missed gene. SNPs didn't help Rosalind Harding find the sickle cell gene.

SNPs seem simple in part because the wealth of genome data being generated by the Human Genome Project and a range of faster, cheaper ways to find SNPs (*Science*, 15 May, p. 1077; 17 July, p. 363) are causing these markers to pile up quickly in both public and private databases. They are much more plentiful than other markers, such as microsatellites, used as genetic landmarks for tracking genes. And they have the added advantage of existing within genes as well as near them, possibly making them useful in identifying the specific variant of the gene that causes disease.

Indeed, most previous gene hunts required studying large, multigenerational families. But in 1996, epidemiologists Neil Risch at Stanford University in California and Kathleen Merikangas at Yale University in New Haven, Connecticut, suggested that SNPs might even be used to track down genes in unrelated people, particularly when the gene merely increases the risk for a disease. This would involve looking for differences in the patterns of SNPs between healthy and unhealthy people (*Science*, 13 September 1996, p. 1516). Prospects such as those led prominent geneticists, such as Francis Collins, director of the National Human Genome Research Institute, and Aravinda Chakravarti of Case Western Reserve University in Cleveland, Ohio, to propose that researchers find enough SNPs to perform such association studies (*Science*, 28 November 1997, p. 1580).



SNP hunt hurdle. Joseph Terwilliger cites perils of genome complexity.

But as two groups reported at the Skokloster meeting, using SNPs to track genes may be less straightforward than thought. Both groups had problems in trying to use patterns of DNA variation to link test genes to diseases with which they were already known to be associated—heart disease in one case and sickle cell anemia in the other.

Working with Charles Sing from the University of Michigan, Ann Arbor, and his colleagues, population geneticist Andrew Clark from Pennsylvania State University in University Park focused on heart disease risk, first examining the role of the lipoprotein lipase (*LPL*) gene. Previous studies had shown that this gene, when mutated, causes high blood lipid concentrations and an increased incidence of heart disease in some families. Clark and his colleagues decided to use SNPs to find out which, if any, *LPL* gene variants might be increasing the risk for heart disease in the general population.

To do this, Sing's team first sequenced a 9700–base pair region of DNA containing the *LPL* gene in samples obtained from 24 people from each of three populations: one in Finland, the second in Rochester, Minnesota, and the third in Jackson, Mississippi. The researchers found that the region contained 88 SNPs, seven of which were in the protein-coding regions of the gene.

Clark and colleagues wanted to use the SNPs in epidemiological studies aimed at understanding the complex chain of genetic and environmental factors that affect heart disease risk. To do this, they first tried to construct a tree representing the historical sequence of mutations that gave rise to the SNPs. The idea was to group different variants of the *LPL* gene according to their ancestral relationships and then compare disease risk among the different lineages. But it immediately became clear that this would be difficult if not impossible; parts of the gene had been shuffled by recombination, the DNA exchanges that occur between the maternal and paternal copies of a gene during sperm and egg formation.

"There has been almost as much recombination as mutation," Clark reported, and that, he adds, "is going to make SNP mapping and association tests much more difficult." Recombination can break down the correlation between the SNPs and variants that inflate disease risk, making it much harder to identify the association. Everybody hopes that only a few regions of the human genome will exhibit this high level of recombination, but where it does occur researchers will need many more SNPs to increase the odds of finding some that correlate with the pertinent mutations.

Rosalind Harding of the John Radcliffe Hospital in Oxford, United Kingdom, shares Clark's concerns about the utility of SNPs, particularly if researchers try to depend on SNPs alone to identify disease genes. A single-base change in the β-globin gene has long been known to cause sickle cell anemia, and she tried to see if SNPs would re-

* The 1st International Meeting on Single-Nucleotide Polymorphism and Complex Genome Analysis was held in Skokloster, Sweden, 29 August to 1 September.

More SNPs on the Way

Late last year, the National Cancer Institute (NCI) launched a project to find genome markers called single-nucleotide polymorphisms, or SNPs, to use in tracking down the hundreds of genes thought to affect cancer risk. NCI has already put about $1 million into the project, called the Genetic Annotation Initiative (GAI), which began generating SNPs in the spring. Researchers running the initiative are hoping that their approach will avoid many of the problems in using SNPs discussed at a recent conference in Skokloster, Sweden (see main text).

NCI is taking what NCI geneticist Ken Buetow, who oversees the GAI project, calls a "gene-based" approach. Instead of creating a genomewide map of anonymous SNPs, Buetow says, NCI will look for SNPs in the coding regions, and in the sequences at both ends, of several thousand genes suspected of contributing to cancer susceptibility or resistance. Besides the 100-plus known cancer-promoting oncogenes and the three dozen or so tumor suppressor genes, the pool will include DNA repair genes, genes that drive the cell division cycle, and genes involved in drug metabolism, immune responses, embryonic development, and cell migration and metastasis. Genes from the NCI's huge Cancer Genome Anatomy Project, which aims for a complete genetic profile of cancer cells (*Science*, 16 May 1997, p. 1023), will also be included as they're identified.

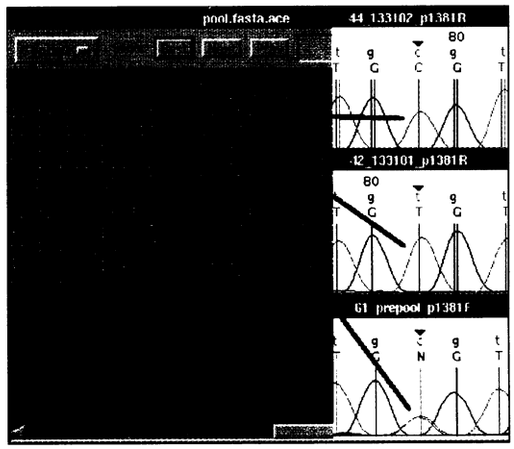
Buetow expects the average gene to yield three to five SNPs, a marker density that makes it much more likely that at least one will be close enough to any cancer mutation to be inherited with it as a block—a phenomenon called linkage disequilibrium. That doesn't ensure researchers won't miss the mutation when screening cancer patients—one of the researchers describing SNP problems at Skokloster had just such an experience with the sickle cell gene—but it should help. "We are less dependent on linkage disequilibrium relationships existing over long distances," says Buetow. "We're going to be right inside the genes." The data generated by GAI will also help determine how common the problems reported at the meeting are.

Once identified, the SNPs will be posted on a database of the National Center for Biotechnology Information, where researchers can access them and design and conduct "association studies" to see if the SNP patterns of cancer patients are different from those of controls. The hoped-for result: hundreds of new cancer genes. Cancer researchers welcome the new initiative. "Given the present technology, it seems to be the obvious next step," says Sofia Merajver, a breast cancer researcher at the University of Michigan, Ann Arbor.

Several issues are still up in the air, however. NCI hasn't decided which populations to screen for SNPs. Right now it's using the DNA of four people from the largely Caucasian families collected at the Centre d'Études du Polymorphisme Humain in France. The GAI wants more diversity, but no one agrees on what that means. "There is concern about stigmatization of populations and concern about what is a representative population," says Buetow. "There are going to be dramatic differences [in SNP frequency] based on geography."

Also under debate is the question of how deep to dig for cancer SNPs. Some would be satisfied with the common ones, in which case screening as few as eight individuals should yield the vast majority. But others argue that the newer, rarer SNPs are also needed, because they're more often in linkage disequilibrium with cancer mutations and thus more likely to show up in cancer association studies.

But the biggest question mark is what technology will be used to discover SNPs and then to detect or "score" them in cancer patients. "Not only has this not been done on a mass scale, but new technologies are being developed so fast, it's hard to know what to do," says the NCI's Mike Dean. To begin, Dean is using a high-performance liquid chromatography mutation-detection method developed by Peter Oefner of Stanford University. Buetow is doing conventional gel-based sequencing, which would be tedious and expensive for large-scale studies.



Gene guide? SNPs, such as the cytosine (C) to thymine (T) change shown here, may point to cancer genes.

One technology now in high demand is the DNA "chip," which can quickly identify SNPs across long stretches of DNA. Affymetrix, a Santa Clara, California, biotech company, has developed such chips, which researchers at the Whitehead Institute for Biomedical Research at the Massachusetts Institute of Technology and Affymetrix are using to do SNP prospecting (*Science*, 15 May, p. 1077). The National Institutes of Health is now negotiating with Affymetrix for a license, and both parties are optimistic. "We would be very happy to collaborate with the NIH in the area of SNP discovery," says Robert Lipshutz, Affymetrix's vice president of corporate development.

Whatever the outcome, Buetow is optimistic about finding methods that will make all kinds of cancer gene discovery projects easy. "We hope to push the technology to enable investigators to do any kind of study they want to do," he says. **–KEN GARBER**

Ken Garber is a science and health writer in Ann Arbor, Michigan.

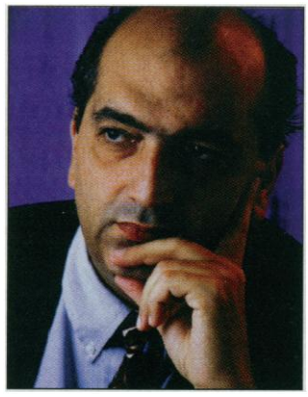
veal the mutant gene. By analyzing DNA samples from 500 people randomly selected from around the world, Harding and her colleagues found that the β-globin gene has dozens of SNPs located in and around its coding sequences.

One of these SNPs turned out to be the sickle cell mutation itself. But when Harding looked at the frequencies of individual SNPs in the 500 samples, and also at inherited SNP patterns called haplotypes, searching for some sign that a particular SNP or haplotype was different, she found nothing that pointed to the sickle cell mutation. With SNP data alone, Harding concluded, "there will be nowhere near enough information to find something unusual and say 'there's a disease gene.'"

She predicts that, in addition to relying on SNPs, researchers will need to know about the patterns of disease and the history of the people being studied. "There has been this naïve idea that once you've gotten to the gene, you'll be able to decide which is the [pertinent] mutation," she adds. "But this is going to be very hard." Others concur. "You can't have just SNPs on their own," says Nigel Spurr, a geneticist with SmithKline Beecham in Harlow, United Kingdom. "You must have [other information and technology] to go with it."

For statistical geneticist Joseph Terwilliger of Columbia University in New York City, Harding's and Clark's experiences with SNPs are indicative of the underappreciated complexity of the genome and of the pitfalls of thinking SNPs will easily

GENSET



Optimist. Genset's Daniel Cohen says SNP problems can be solved.

lead geneticists to elusive disease genes. "Risch and Merikangas have been taken out of context" by overly enthusiastic promoters of SNPs' potential, he argues.

Terwilliger notes that although Risch and Merikangas found association studies practical for identifying disease genes in which one mutation accounts for most of the increased risk, that situation may be uncommon. In a survey of all the new disease genes reported in the *American Journal of Human Genetics* during the past 1.5 years, Terwilliger found that about 90% of those genes had more than 10 pertinent mutations that predispose an individual to disease. With so many different mutations involved, none is likely to stand out in a SNP analysis. And that's the easy case, involving diseases caused by mutations in a single gene. The situation will be worse for cancer and the many other diseases in which multiple genes contribute to increased risk. "It's not just the underestimated complexity of the genome as much as it is the underestimated complexity of the etiology of a complex disease," he adds.

Researchers at Skokloster agreed that it will be difficult to gauge the usefulness of SNPs until they know more about how genomes vary between and within the world's ethnic groups. Because the most universal SNPs will be among the oldest, they are likely to exist in people both with and without disease. This means that there may be no distinctive pattern of SNPs specifically associated with a key variant of a gene. It could be easy to miss an important association or to make an association with the wrong gene variant. "If we don't think carefully before we do these experiments, we'll wind up with a lot of false signals," Uppsala's Brookes says.

Others at the meeting pointed out that association studies require that researchers look at much larger numbers of people than typical family studies, to sift out the false signals. "It's not enough to have 70 controls and 50 patients," says Gert-Jan Van Ommen, head of the Human Genome Organization and a geneticist at Sylvius Laboratories in Lieden, the Netherlands. "You're talking about requiring populations of several thousand." SNP analysis won't begin to be useful without new, high-speed technology for analyzing the thousands of DNA samples required, says Spurr.

Even with these caveats, however, the researchers expect to see SNPs research proceed. "We know they will be successful in certain situations," comments Case Western's Chakravarti. "We just don't know how successful they will be."

Already, association studies have linked a few gene variants to diseases. The tying of the *ApoE4* gene to an increased risk for Alzheimer's disease in Caucasians is one often-cited example. And geneticist Daniel Cohen, head of Genset in Evry, France, says that his company has worked out many of the issues raised by the conference participants, in part by developing new methods—which he would not describe in detail—for analyzing the data and discerning real associations. In October, for example, he plans to announce the SNP-based discovery of two genes involved in prostate cancer. "I am absolutely confident of this strategy," says Cohen. "It works."

Although others may not share Cohen's confidence, they want SNPs to be put to work. "Provided they are not being regarded as the panacea for complex disease findings, there is value in producing SNPs," says Van Ommen. "[SNPs] are going to make a big difference." **–ELIZABETH PENNISI**

ECOLOGY

Software Helps Australia Manage Forest Debate

A computer program to promote biodiversity gives loggers and conservationists a chance to end their fierce fighting over forest reserves

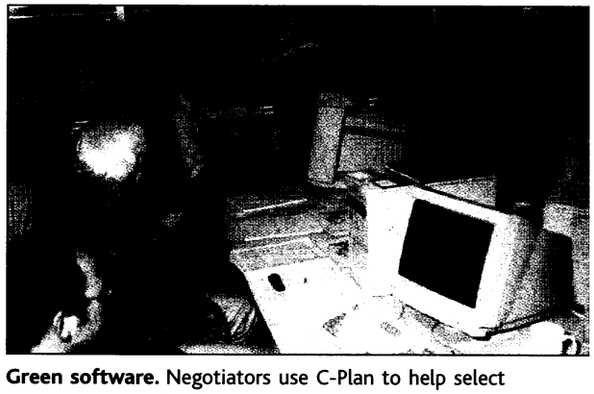
MELBOURNE, AUSTRALIA—The forests of New South Wales (NSW) have seen many bitter battles in the last 20 years between logging interests eager to feed an insatiable Japanese appetite for wood pulp and conservationists trying to preserve the country's dwindling arboreal heritage. Those battles have taken a heavy toll on the participants. Just ask Col Dorber, the executive director of NSW Forest Products Association. In 1995, Dorber suffered a stress-induced heart attack after being roundly condemned by government and industry officials and vilified in the media for publicly defending a logger caught punching a "greenie."

Now back on the job, Dorber sees his remarks as an unfortunate reflection of the historic enmity between the two camps. That's why he's so encouraged by an experiment drawing attention from ecologists and resource managers around the world that attempts to inject science into forest management and that respects the interests of all parties. "Since 1995, we've been through a culture change," he says. "Prior to that, we [industry and conservationists] wouldn't speak to each other. But now we've learnt to respect each other. It's a fantastic process."

That process is a joint initiative by the federal and state governments to negotiate long-term agreements for forest reserves that allow continued logging while maximizing biodiversity. At the core of the negotiations is a computer program, called C-Plan, that gives adversaries a chance to trade in their swords for software. Like some ecological card game, the software puts a biodiversity value on each parcel of land and presents stakeholders with various packages that meet the conservation target. C-Plan was developed by NSW National Parks and Wildlife Service conservation planners Bob Pressey, Simon Ferrier, and colleagues, and programmer Mathew Watts at the University of New England in Armidale, NSW. So far it has been used in two major sets of negotiations; a third exercise, involving a large swathe of old-growth forest, has just begun.

"It's setting the gold standard in the field," says ecologist Reed Noss, co-executive director of the Conservation Biology Institute in Corvallis, Oregon, and president of the international Society for Conservation Biology. Indeed, the World Bank is using C-Plan for an assessment in Guyana, and Pressey is currently in South Africa to help

TOM BARRETT



Green software. Negotiators use C-Plan to help select reserve areas in Eden forest.