

A Summary of the  
**SYMPOSIUM ON**  
***THE ENVIRONMENTAL***  
***GENOME PROJECT***

**October 17 – 18, 1997**

*Sponsored by*

National Institute of  
Environmental Health Sciences (NIEHS)  
National Institutes of Health (NIH)

## CONTENTS

Contents .....	iii
Foreword .....	v
Agenda .....	vii
Overview .....	xi
Session I: Gene-Environment Interactions in Human Diseases .....	1
Introduction .....	1
Examples of Polymorphisms in Disease Susceptibility .....	3
Examples of Gene-Environmental Interactions .....	4
Issues Associated with the Environmental Genome Project .....	7
Session II: Population Sampling .....	9
Introduction .....	9
Allele Identification .....	10
Frequency Distributions of Polymorphisms .....	12
Session III: Technologies .....	15
Introduction .....	15
DNA Sequencing .....	16
Chip Technology .....	17
Session IV: Population-Based Epidemiological Studies .....	19
Introduction .....	19
Opportunities .....	21
Optimizing Study Designs and Power .....	24
Session V: Functional Analysis of Polymorphisms .....	27
Introduction .....	27
Functional Analysis Models .....	29
Gene Expression Assays .....	31
Session VI: Ethical, Legal, Social Issues .....	33
Introduction .....	33
Ethical and Social Issues in Sampling .....	34
Informed Consent, Potential Impact on Susceptible Groups .....	38
Program Participants .....	41

## FOREWORD

In October of 1997, the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), on its Bethesda campus, sponsored a symposium, a two-day series of talks to help shape a comprehensive research initiative: the Environmental Genome Project. This special report will provide attendees a refresher course on the issues raised by our distinguished panel of speakers, among the top experts in their disciplines. We hope it will give other readers who were not present, whether they be scientists, policymakers or the lay public, a sense of the broad, complex terrain covered during the symposium. We hope, too, this report will serve as a primer on the themes that are likely to inform and guide the Environmental Genome Project for years to come.

Kenneth Olden, Ph.D.  
Director

## AGENDA

**Friday, October 17**

- 7:30 a.m. Registration
- 8:15 a.m. *Welcome and Overview Remarks*  
Kenneth Olden  
Samuel Wilson
- 8:30 a.m. **Session I - *Gene-Environment Interactions in Human Diseases***  
J. Carl Barrett, Chair
- Introduction*  
J. Carl Barrett and Leland H. Hartwell
- Topics:** *Examples of Polymorphisms in Disease Susceptibility*  
*Examples of Gene-Environment Interactions*  
*Issues Associated with the Environmental Genome Project*
- Speakers: Douglas A. Bell  
Jack A. Taylor  
J. Carl Barrett
- General Discussion*
- 10:00 a.m. COFFEE BREAK
- 10:30 a.m. **Session II - *Population Sampling***  
Dean H. Hamer, Chair
- Topics:** *Allele Identification*  
*Frequency Distributions of Polymorphisms*
- Speakers: Kenneth M. Weiss  
Charles H. Langley
- Discussants: Georgia M. Dunston  
Kenneth H. Buetow
- 12:00 p.m. LUNCH

1:30 p.m.

**Session III - Technologies**  
Glen A. Evans, Chair

**Topics:** *DNA Sequencing*  
*Chip Technology*

Speakers: David G. Wang  
David R. Cox

Discussants: Deborah A. Nickerson  
Maureen T. Cronin  
Peter Oefner

3:00 p.m.

COFFEE BREAK

3:30 p.m.

**Session IV - Population-Based Epidemiological Studies**  
Stephanie J. London, Chair

**Topics:** *Opportunities*  
*Optimizing Study Designs and Power*

Speakers: John D. Potter  
Duncan C. Thomas

Discussants: Susan E. Hankinson  
Joseph F. Fraumeni

6:00 - 7:30 p.m. **RECEPTION** - Bethesda Marriott Hotel  
Congressional Ballroom  
5151 Pooks Hill Road

**Saturday, October 18**

8:30 a.m.      **Session V - *Functional Analysis of Polymorphisms***  
Samuel H. Wilson, Chair

**Topics:** *Functional Analysis*  
*Models*  
*Gene Expression Assays*

Speakers: Frederick P. Guengerich  
Patrick O. Brown

Discussants: Leland H. Hartwell  
Edison T. Liu

10:00 a.m.                      COFFEE BREAK

10:30 a.m.      **Session VI - *Ethical, Legal, Social Issues***  
Kenneth Olden, Chair

**Topics:** *Informed Consent*  
*Ethical/Social Issues in Sampling*  
*Potential Impact on Susceptible Subgroups*

Speakers: David J. Hunter  
Francis S. Collins

Discussants: David C. Christiani  
David R. Cox

12:00 p.m.                      **CLOSING REMARKS**

## OVERVIEW

The Environmental Genome Project is best introduced as a work in progress, an outgrowth of the longstanding interest of NIEHS and its grant recipients in relationships between environmental exposures, diseases and their genetic influences. Everything about this new research program is open to discussion. As with any evolving science, the Environmental Genome Project will continue to be molded by new ideas and approaches.

We begin with the premise that, as never before, genome science offers many opportunities for improving public health. At the same time, it offers many challenges. Some—data collection and interpretation—are scientific. Others are social, legal and ethical—for instance, how do we go about collecting that information and what implications does knowledge about genetic susceptibility and environmental exposure have beyond science? Still others are pragmatic, about framing these questions in ways that can be addressed within budget limitations. And still other challenges involve using information about genetic susceptibility to design and conduct studies that will eventually reduce the burden of environmental exposure-related disease.

## Session I: Gene-Environment Interactions in Human Diseases

*“We are all at risk for something.”*  
– Francis Collins

### Introduction

Why is it that so many people get sick and die before their time? Research into that question has zeroed in on environmental circumstance and genetic predisposition.

That harmful agents in the environment can cause diseases would seem self-evident enough—and the solution equally so. Identify what in the environment causes a disease, then take measures to prevent exposure to it.

But even such a seemingly simple notion as exposure is not so simple. “Environmental exposure” can be defined broadly. It includes exposure not only to natural and synthetic chemicals but also to radiation, substances in food and even an individual’s nutritional status. The definition also encompasses social and economic status, which seem to have a great though imprecise impact on health.

Cases of an environmental contribution to disease are abundant. Toxicologists have identified many agents in environment-linked diseases, many tied to the workplace. It is a much more daunting task to show how exposure to minute traces of chemicals might degrade health of a large, diverse population, especially if those traces are ubiquitous.

Strong genetic predisposition to environmentally-induced disease has been observed, and in several cases, such as cystic fibrosis and breast cancer, it is partly understood. However, the hereditary forms of those diseases are rare. It appears that disease-implicated genes in different individuals can lead to very different effects, depending on environmental exposure. Studies to be carried out under the umbrella of the Environmental Genome Project propose to explore genetic susceptibility to environmentally-associated disease.

Which diseases in particular? Environmentally-associated diseases seem to know no boundaries. The same agent can play a role in cancer and birth defects, reproductive failures and immune disorders and on and on. So understanding fundamental gene-environment interactions can lead to understanding susceptibility to many diseases.

The key to understanding how these agents lead to disease in diverse populations, and to gauging susceptibility in communities at large, may lie in common genetic differences, or polymorphisms. A polymorphism is a genetic variation that exists in more than 1 percent of the population. Therefore, an important goal will be to identify functionally important variations in DNA sequence, common polymorphisms, in known genes that are likely to be influenced by environmental exposures. Epidemiologists can then use this information to identify susceptible subpopulations, ultimately enabling them to determine low-level

exposures that affect health. These data can be crucial for preventing environmentally-associated diseases.

It will be important to distinguish between genes that act predominantly on their own to cause disease from those which act in concert with environmental agents; these genes may not in and of themselves put an individual at risk but may put someone in a given environment at risk. Such information can help decrease the probability of disease for affected individuals by limiting their exposure to the gene-interacting agent.

Many laboratories are already at work identifying environmental-response polymorphisms. As labs participating in the Environmental Genome Project contribute to this goal in the future, the rate of progress will increase. Other areas of interest for the Environmental Genome Project include epidemiological studies of gene-environment interaction, and investigating new tools for genetic analysis. The Project will look at finding new ways to design studies and build sample repositories. It will also be necessary to remain mindful of a maze of ethical, legal and social implications when dealing with people's varying degrees of genetic susceptibility.

There are two general targets for environmental agents within cells and within tissues. Agents can upset genetic stability directly or perturb the process by which genes repair themselves. Either way, the processes can lead to disease—or to other disruptions that can cause disease. For example, many agents act like the hormones that activate genes, mimicking signal transduction but with dire consequences to cells.

The polymorphisms likely to be important will be diverse in determining risk to disease. There's a complex interplay between DNA repair, cell-cycle control, cell death and possibly differentiation. Genetic variation can lead to differences in responses to environmental exposures. Identifying groups genetically susceptible to disease makes early intervention possible.

For example, a Western Washington study showed that when people predisposed to esophageal cancer were monitored yearly and received surgery at the first sign of cancer, their chances of survival were much greater than had they not been closely watched.

Although the work proposed by the Environmental Genome Project offers great promise, there should be caution concerning its potential as well. An analogy could be drawn that might help us assess our ability to formulate correctly, let alone answer, the questions that the Environmental Genome Project seeks to answer. The analogy draws on the field of enzyme structure-function relationships. Thus, at our current state of knowledge of the human genome we are like the enzymologist looking at a low resolution protein structure, one that only shows the folding pattern of the polypeptide. Ten years from now we would be like the enzymologist looking at a high resolution structure that reveals the exact position of all amino acid side-chains in the enzyme active site. Today, we could not begin to approach the question of structure-function. But 10 years from now we would be well-equipped to do so. In 10 years, all common polymorphisms in the human genome

may be identified. In the meantime, while only a fraction of the information relevant to our questions is available, how can something useful be undertaken that won't be immediately supplanted when new information becomes available?

One way to view the Environmental Genome Project would be to assume that we know a lot. The other view assumes that not much is known. For instance, a typical gene from the human genome has a small coding sequence that might have one or more common variations and will be surrounded by DNA with many more polymorphisms. What is the important variation in this single gene? In most cases, no one knows.

There are hundreds of thousands of genes in the human genome yet a scientific approach that concentrates on one gene at a time, a single-gene orientation, is prevalent and persists. As technology allows approaches to study many genes at the same time, the view and vocabulary will change, from single genes to sets and patterns of genes, patterns of polymorphisms and of expression.

Several questions that confront us will guide the initial design of the project. We must decide how many genes should be examined. Which genes? How much sequence from each gene, from coding regions versus non-coding? How many alleles does one look at for common polymorphisms? What populations should be studied? In attempting to answer these many questions and shape the form of the Environmental Genome Project, one important challenge should be in the thoughts of all participants: that challenge is the need to choose and achieve a balance between the quantity and type of the work of the Environmental Genome Project.

### **Examples of Polymorphisms in Disease Susceptibility**

If you look at three similar faces, you can tell they have a common genetic background that accounts for their sameness. You might look at faces with similar eyes, nose and smile and figure that they are siblings. But when you look at these traits in people around the world, diversity is the rule. It is the same for other kinds of genetic variation, including some that are very rare and cause disease. But there are other more common variations like blood type and alcohol intolerance that are considered to be of polymorphic frequency.

If you think about the enzymes that ultimately determine these traits and look at the distribution of enzyme activity in a population, you notice that some enzymes have a very narrow range of activity. People who inherit alleles outside this enzyme-activity range generally suffer from a genetic disease and face unfavorable survival odds.

Other enzyme activity distributions are broader. In some cases, low-enzyme-activity alleles are rare across populations, high-activity alleles more common. Or vice versa. Some alleles are frequent enough to be polymorphic in frequency but may or may not have functional significance. Polymorphisms may affect functions if the variation occurs within the coding structure. It may occur upstream, in the regulatory region, in the on-off

switches for the gene. In some cases, an entire gene may be deleted; in others, genes are duplicated, perhaps many times.

How common are polymorphisms in these genes? A polymorphism that is present in one population generally occurs in another at some, usually different, frequency. Some common polymorphisms in one group are rare in another. It should be noted that for a variation to reach 1 percent frequency and fit the definition of a polymorphism, a mutation must have occurred thousands of years ago. In such a case, the polymorphism could be explained as a random effect of migration and isolation. Polymorphisms at high frequency may also be an effect of natural selection. In sickle cell anemia, for example, people heterozygous for the disease gene also have a survival advantage in areas where malaria is prevalent. Similarly, polymorphisms in xenobiotic metabolism enzymes may have been advantageous to early humans trying to utilize toxic food sources.

How can polymorphisms affect health? For one, polymorphisms in xenobiotic metabolism genes can cause adverse reactions to drugs. Polymorphisms can also modify the response to the environment in exposure-associated disease—in cases of exposure, people with susceptibility alleles will be at higher risk than those with resistance alleles. The converse is also true and can be seen with HIV exposure, in which individuals with the CCR5 deletion allele have been observed to possess a greater latency in the development of AIDS than those with the more common sensitive allele. This variation is probably of little significance in the absence of HIV exposure. Similar observations have been reported in the polymorphic alleles of the GSTM1 gene with respect to bladder cancer risk and smoking. Smokers with the resistant or null GSTM1 allele have no increased risk compared to people with other variants.

Alcohol metabolism is another interesting example. The metabolites of alcohol have unpleasant effects many hours after consumption. Some people carry no functional alleles or low-activity alleles for aldehyde dehydrogenase. In those people, drinking alcohol produces very high levels of acetaldehyde in their blood, manifested by flushed skin and extreme discomfort. As you might expect, these people don't enjoy consuming alcohol and, it turns out, that homozygotes for this gene don't become alcoholics and are at low risk for alcohol-related illnesses including liver, mouth and esophageal cancer, and cirrhosis of the liver. The point: genetics is actually influencing drinking behavior, a dietary choice. In a curious twist, some heterozygotes are able to tolerate high levels of acetaldehyde in their blood, overcoming the incumbent unpleasantness, and become regular drinkers, even alcoholics. In a study from Japan, regular drinkers in the heterozygote genotype were found to be at a much higher risk for alcohol-related cancers.

### **Examples of Gene-Environment Interactions**

There are many classic gene-environment interactions that cause disease. Phenylketonuria, for instance, is an inability to convert phenylalanine to tyrosine. Children who inherit this recessive trait develop high blood phenylalanine, which can lead to mental retardation. A simple test ensures that all infants are screened for this

disease, which can be treated by restricting phenylalanine in the diet. UV radiation is another example of how genotype increases exposure-associated risk. UV radiation, well-known to cause skin cancer, triggers mutations primarily at dipyrimidine sites. Genetic predisposition to skin cancer is caused by genetic alterations in genes that repair UV-induced DNA damage. This occurs in the human disease Xeroderma Pigmentosum.

The Environmental Genome Project does not intend to concern itself with these particular classic examples. Rather, it intends to consider common polymorphisms in the population, variants that occur at more than 1 percent. Some examples of these, among the cases in the literature are: beryllium and HLA in chronic beryllium disease, or CBD; maternal smoking and TGF- $\alpha$  in facial clefts; and, finally, arylamine metabolism and NAT1 and NAT2 in bladder cancer. These high-risk alleles, heterozygous or homozygous, are among many that occur in an appreciable portion of the population. And the genotype and exposures—heavy metals, diet, viruses—lead to a variety of diseases, from lead poisoning to myocardial infarction. Some of these variations are protective, such as CCR5. As reported in the previous section, an individual allele may put you at increased risk for one disease and a decreased risk for another, so you have to be careful about labeling particular variants as bad or good.

Beryllium is a modern industrial exposure affecting workers in ceramics and electronics and, in particular, nuclear weapons. About 2 to 5 percent of exposed workers (as high as 16 percent in a few cases) develop CBD, characterized by lung granulomas invaded by a particular form of T cell. If you take the lymphocytes from one CBD patient and expose them to beryllium salts, the lymphocytes will proliferate. This is the basis of a CBD screening test. If you take T cells from CBD patient lung granulomas and clone them, they respond to the beryllium salts only if the antigen-presenting cell, the one that presents the beryllium to the T cell, expresses one of the major histocompatibility complex, or MHC, class 2 proteins on its cell surface. MHC 2 genes are HLA genes. Allelic HLA variants have been implicated in susceptibility to several autoimmune disorders, which suggests that the HLA genotype affects susceptibility to CBD. A study published several years ago in *Science* looked at a particular variant of the HLA DPB1 gene. That particular allelic variant was present in about 30 percent of the healthy unexposed population. Workers exposed to beryllium but disease-free also showed this allelic variant at about a 30 percent frequency. But in people who developed CBD, 97 percent had this particular allelic variant. This is a straightforward example of a gene-environment interaction. You must be exposed to beryllium to get the disease, but it also appears that the susceptibility genotype for this HLA is a critical factor in who develops CBD.

Facial clefting is a common birth defect that results from incomplete closure of the palate. Transforming growth factor- $\alpha$  (TGF- $\alpha$ ) is highly expressed in palate cells as the two sides of the palate merge. Evidence indicates a polymorphic TGF- $\alpha$  allele increases the risk of clefting in newborns and is associated with maternal smoking. Mothers with the uncommon A2 TGF- $\alpha$  allele had a slightly but statistically insignificant increase in risk of

clefting in the newborn; however, mothers with that allele who were heavy smokers had a roughly two-fold elevated risk of having a child with clefts. This gene-environment interaction is an example of a growth factor related locus that increases exposure-associated developmental risks. There was some increased risk associated with the variants alone but it was statistically insignificant. When smoking was thrown into the equation, there was an elevated relative risk—another example of the gene-environment interaction.

Bladder cancer is a common disease, about fifth in rank of cancer incidence in the U.S. population. Increased risk from smoking ranges from two- to 10-fold; some, but not much, is of familial predisposition. Ethnic differences persist even after adjusting for smoking, and a number of environmental and industrial exposures increase the risk of developing bladder cancer—in particular arylamine exposure. Industrial exposure to arylamines have shown incredibly increased risk. Twenty- to 40- to 100-fold increased risk in some cases. Arylamines are constituents of tobacco smoke and may increase the risk of developing bladder cancer in smokers.

There are two key players, predominantly expressed in the liver: NAT1, in which a particular allele, NAT1-10, is a putative fast-metabolizer versus the other wild-type versions; and NAT2, a polymorphic gene with "slow" and "fast" genotypes. A study by Jack Taylor, NIEHS, showed that arylamines can reach the bladder via urine and can be subsequently taken up by the bladder, where NAT1, highly expressed in the bladder, can acylate it to create a highly reactive compound that can cause DNA adducts and lead to cancer. The study looked at NAT1 and NAT2, using smoking as a surrogate for arylamine exposure. Genotypes aside, the study showed that smokers incurred a roughly three-fold increased risk for developing bladder cancer. Looking at NAT2 genotypes, slow versus fast, and ignoring smoking, there were no increased risks, no suggestion that NAT2 played a role in bladder cancer. Looking at the NAT1-10 allele, there was a hint that the NAT1-fast genotype might increase the risk of bladder cancer. When smoking and genotype were linked, nonsmokers, regardless of genotype, had a low cancer risk. But among smokers, the risk jumped; those who carried the NAT1-10 allele had a higher risk than smokers who do not carry that allele—a clear gene-environment interaction. Further, homozygotes for this allele had a higher risk for developing bladder cancer than homozygotes for wild-type with increasing years of smoking.

While the study showed that NAT1 had an effect on developing bladder cancer and NAT2 alone did not, another study led by Doug Bell, NIEHS, showed that people who were slow for NAT2 and carried the NAT1-10 allele had more adducts in bladder epithelial cells compared to other genotypes. A subsequent Taylor study of NAT2 and NAT1 together showed that in nonsmokers, genotype had little to do with one's risk of developing bladder cancer. However, among smokers who were NAT2-fast, there was a jump in risk that was not present in the NAT1-10 genotype; in people who were NAT2-slow, there was an appreciable effect in the NAT1-10 genotype. Why? That is still being sorted out, but this serves as an example of a three-way interaction: gene-gene-environment.

These and other studies show that many common polymorphisms are at low penetrance but have high attributable risk if the exposure related to them is common, and these polymorphisms and exposures can explain many diseases in the population. Such gene-environment interactions may help identify the mechanisms by which exposure causes disease. A big challenge of the Environmental Genome Project will be to look at those mechanisms in detail. The simultaneous study of genes and environment and of gene-environment interactions will be important both for identifying the alleles of genes that cause disease, the exposures that cause disease and, in turn, to help identify true risks associated with environmental exposures.

### **Issues Associated with the Environmental Genome Project**

The specific goals guiding the Environmental Genome Project are as follows:

To establish a repository of DNA sequences that represent the U.S. population, which will enable researchers to search for polymorphisms and will provide them with a resource for studies now and in the future.

To catalog polymorphisms based upon the resequencing of specific genes selected for study in this project.

To devise a means for determining which polymorphisms are functionally important in human disease.

And to facilitate epidemiological studies of gene-environment interactions.

A project with goals this ambitious raises many issues. The first issue is the criteria for selecting genes to study. This will be the responsibility of a committee, to convene following this symposium. In part, the purpose of this symposium is to get as broad an input as possible on gene selection. NIEHS is soliciting nominations for genes to study from individuals and organizations, in a variety of ways, including a web site. Anyone can nominate a gene to be considered by this working committee [<http://www.niehs.nih.gov/envgenom/genenoms.htm>].

There are approximately 100,000 genes in the human genome; but this will not be a gene-discovery project. In fact, this project will exploit information provided by the U.S. Department of Energy and the Human Genome Institute's gene-identification effort. Once the candidate genes are identified, more issues will arise. For example, which genes will have first priority? The number of genes considered will be narrowed automatically by selecting those whose sequences are already known. Preference will be given to genes not only whose sequences are known but, more importantly, to those implicated for their role in environmentally-associated diseases. Since not much is known about gene structure, function and interactions, any information that exists in this area might be another mechanism that will drive gene-selection.

After gene selection, the majority of this project will involve resequencing environmentally-associated genes to establish a catalog of polymorphisms. The issue then becomes which technology will work best for this resequencing phase. Should the Project look at cDNA, which cover the coding sequences? Or should genomic DNA also be studied? There is also the issue of whether to look for polymorphisms in non-coding regulatory regions? This may be crucial. And there is the question of how one determines the functional significance of a polymorphism once it is discovered.

The Project will identify polymorphisms by first resequencing the coding regions, the exons, from the genomic DNA. Next, the Project may derive sequences of putative regulatory regions in mouse and human non-coding regions and compare them. There will also be some sequencing of cDNA, seen as an easy way to gain a lot of information quickly from large numbers of subjects. And the Project may sponsor sequencing to fill in gaps in information about variation between individuals and to complete genomic structures of important genes.

What will all this cost? There are many variables that figure into the cost of the resequencing phase, no less the functional analysis and the epidemiological studies. There are many estimates of what it costs to sequence genes today, ranging from a few pennies per base to 50 cents per base. The best estimates today put resequencing of a gene at around 10 cents per base, which is why this project will focus on resequencing rather than gene discovery. The hope is that with new technology such as chip-based resequencing, the cost can be reduced to a few cents per base. The consensus is that the cost to resequence in the first half of this project will be far greater than that associated with the second half. This is because technology will advance over the course of this project. Today, at 10 cents a base, the project will process  $10^4$  base pairs per gene. This, it can be argued, may be too many in some cases, too few in others, but it's a reasonable estimate. The goal is 200 genes or more. A representative sample of the population requires 500 to 1,000 individuals—or \$200 million, considered too large a price tag for this project.

There are many options to reduce costs. One way to cut resequencing costs would be through chip technologies, but developing such technologies is not part of this project. Chip technology is, however, being actively pursued by the Human Genome Institute, and this project will be a beneficiary when the technology becomes available in the next few years. This project also could consider sampling fewer individuals. Are 1,000 subjects necessary for a representative sampling, or can there be fewer? Fewer genes is an option, too, but a last one. Perhaps limiting the project to cDNAs is an option, but important information might be lost, and it is not an optimal alternative, either. The ultimate answer is “all of the above,” to varying degrees.

The next phase will be to determine which polymorphisms are functionally important in environmentally-associated diseases. If genes whose various activities are already known are chosen, it should be possible to identify important polymorphisms that contribute to

these different activities. This has been done for the p450s and other enzymes already. The project could emphasize differences in expression level presumably being controlled by regulatory regions; that would point to where to look for functionally important polymorphisms in those genes. Sequence domains, motifs within a gene, could be conserved, and that too could yield clues about which polymorphisms might be important. Perhaps understanding a gene's three-dimensional structure will help to pinpoint regions of interactions between proteins and environmental agents. Studying gene function in model organisms also will be important, as will other systems in which protein interactions and protein-chemical interactions can be analyzed.

A different approach to understanding functionally important polymorphisms will come from the epidemiologists. When a catalog of polymorphisms in a wide variety of genes are put on a chip that contains a host of mutations, the epidemiologist can conduct large population-based studies and link diseases to functionally important changes.

Indeed, the ultimate goal will be to use this information in epidemiological studies, and that will beget many other issues, such as improving the technologies for genetic analysis and population-based studies, optimizing study designs and developing population-based resources. If researchers can develop ways to share these resources once they devise them, it will be all to the great benefit of the scientific community.

These and other crucial and complex issues, such as the ethical, legal and social implications associated with this sort of information, will be addressed extensively in the sessions that follow.

## **Session II: Population Sampling**

### **Introduction**

There's a cautionary tale in population sampling known as "The Chopstick Problem." Suppose an NIH institute wants to seek out a gene linked to chopstick use, a measurable phenotype. The phenotype varies between people and can be looked at in many ways. Someone conducting a survey of a sample population—the symposium audience, for example—to find out who uses chopsticks, how often and how well, takes a blood sample and looks at different genes, say 200. That investigator would find highly significant associations. Further, other researchers in other places could replicate that experiment and find exactly the same result. The discovery might be called the Successful Utilization of Hand Instruments gene, or SUSHI gene. Perhaps a *Nature* article would ensue, and of course it would be wrong. Such a sampling would have actually turned up various blood group genes or HLA antigens that just happened to differ between Caucasians and Asians, who happen to differ in chopstick dexterity for completely cultural reasons. This is precisely the reason that not only identifying alleles but also looking at their frequency in different populations is so important.

## Allele Identification

A collaborative, multi-institutional project, sponsored by the National Heart, Lung and Blood Institute, focused on variation in a candidate heart disease gene among 72 individuals divided evenly among three different populations: African-Americans in Jackson, Mississippi; North Karelians, people from Eastern Finland, in whom variation is supposedly reduced because of their recent settlement by a relatively small founder population; and a mixed European population in Rochester, Minnesota. Partial data was presented for 24 Mayan-Amerindians from the Yucatan Peninsula.

The project raised a range of questions, including: How much variation is there? How well organized is the variation in terms of its linkage arrangement, in terms of its evolution of chromosomes? Why are epidemiological studies of markers so often inconsistent? For example, a marker association will be found in one population and the same association not found in another population. Even the same marker found in one population study will not be found by another study of the same population.

Kenneth Weiss, Pennsylvania State University, led the study, which examined allelic variation in a 10kb region constituting a quarter of the LPL gene, which is involved in lipid metabolism. They examined 88 variable sites, all but one being diallelic, and found tremendous variability, not only between population groups but also within groups, including the supposedly less variant North Karelians. Weiss noted that this high level of variability was found in a small number of individuals in only one quarter of a single, admittedly large, gene; but 10kb is only 1/300,000th of a human genome.

In Mississippi, 79 of the 88 sites varied; in Finland 56, reduced in variation compared to African-Americans but similar to Minnesota's generalized European population, which had 60 varied sites. A measure of variation called nucleotide diversity--basically the average probability of finding a heterozygote at a given site in the 10kb--was fairly similar in all the populations. Though slightly reduced in North Karelia, nucleotide diversity was about .002 on average, or roughly twice previous estimates for nucleotide diversity in humans.

The frequency of the rarest allele across the 88 sites was distributed fairly uniformly. There was no real evidence of a clustering of regions, in which some set of alleles would have jointly risen to high frequency. Many variants had quite substantial frequency. There were some singletons and doubletons in the data and a lot of variation in different levels of frequency. The frequency distribution of variant alleles was in line with theoretical expectations, based on a model population at genetic equilibrium between mutation and genetic drift, though the data showed an excess of alleles of intermediate frequencies compared to the model.

The ranked heterozygosity, or variability per nucleotide site, was shown for the 50 most variable sites. The Mississippi sites were slightly more variable than those in the Rochester population. The North Karelians were, indeed, less variable but not very

different from the other populations. It was noted that the variability in the Karelian population, which had been thought quite favorable for mapping disease genes, raised questions about the suitability of a study population. Among the 72 individuals, only two individuals varied in only one site within the 10kb and no individual was a complete homozygote. At the other extreme, two individuals varied simultaneously at 39 sites. The average individual was heterozygous at 17 sites. The results showed a great genetic variation among human beings.

Weiss said that 54 of the 88 variable sites were found in all three of the populations. Two variants were found only in the North Karelians. Five variants were found only in the Rochester sample of mixed Europeans. One variable site was found only in the two European samples. A comparatively large, 18 variable sites were found only in the Jackson sample, consistent with the previous discovery that African-Americans and Africans are more genetically variable than other world populations. The data pointed out the challenge of identifying variation adequately because it is so abundant and not uniformly distributed across populations.

Another question of concern for those attempting to link genetic variations to inherited diseases is linkage among alleles. Due to past recombination or mutation, genetic variations may be distributed within a chromosome in various linked arrangements yet still act as one allele. How structured are the variants or 88 sites in terms of chromosomes? One test is how often one finds all four haplotype phases in a sample of this size. Meaning if one finds a big A and a little a at one site and a big B and little b as the two alleles at the other site, how often does one find AB, Ab, ab, and aB in the same data set? There was abundant evidence of past recombination and mutation in sites scattered all over the 10kb region under study.

In a theoretical model, the pattern of variation in a chromosome was subject to the combined influences of recombination and mutation, creating variations at roughly the same rate, compounding one another's effects and complicating chromosome structure. Although the number of alleles reached a plateau as sample size increased, the number of haplotypes continued to increase linearly. In studying 200 chromosomes, the group had not begun to reach a plateau at which they had seen all the haplotypes present in the human population. Haplotype variation was not evenly distributed in the sample--the amount of variation in half the sample would not allow a prediction of haplotype variation in the other half. The 88 haplotypes from the study data (not including the Mayan data) sorted into a tree, based on similarity. This technique revealed a pattern of similarity among haplotypes in specific regions, suggesting the possibility of determining a historical hierarchy of recombination. Haplotype data from 25 of the sites, selected from the original 88 sites for high haplotype structure, revealed two large clades of similar haplotypes.

Unfortunately, the expectation that one clade of haplotypes might be common in one population, say Rochester, while another clade was common in Jackson, was not realized. The number of haplotypes from all three populations fitted into each of the clades in

roughly the same proportion. So though haplotype structure seems apparent, interpreting its meaning is not yet straightforward.

To explain why epidemiological studies that only use a single or two markers are so inconsistent, an association study assigned one of the variable sites in the data as dominant and causative of disease and another as recessive and not causative but potentially useful as a co-dominant marker site. Simulating a search through the 10kb sample, one would be able to detect an existing causal site only about half the time. However, by using the 25 markers that form the two major clades, and assuming the causal site was one determining the clade, detection of the causal site would go up to 82 percent—not perfect but better. But if the causal site was not in the clade, and the cladistic subset is used to attempt to detect the site, detection would drop to 48 percent. Also, if the causative site were in the clade, but one looked at all the other sites, ignoring cladistic structure, detection would drop to 60 percent.

The studies showed that variability among individuals is abundant, and though cladistic structure in data offers hope of making causal inference in some cases, it should not be assumed that screening will detect disease causal sites.

### **Frequency Distributions of Polymorphisms**

Experiments in *Drosophila* in the 1980's raise questions about the ability to address polymorphism in the context of the Human Genome Project and pose a challenge to the ability to survey genetic variation and perform association studies in the human population. At what frequency can we expect DNA sequence polymorphisms?

Theory predicts that the frequency of a deleterious mutation would be increased by an increase in the mutation rate and reduced by selection factors that reduce the mutant's fitness in the population. Genes with a very low mutation rate or very strong negative effects on fitness will be rare, while ones with minor or no negative effects may be much more frequent, depending on the mutation rate. In the case of molecular polymorphism, newly arising mutants or mutants in a population in equilibrium do not have any affect on fitness and are affected only by genetic drift. They also have a rather skewed distribution, with a great preponderance of rare variants. This raises the questions: How rare? And how much variation will there be?

In *Drosophila* it was discovered that in genes in regions where crossing over is very low relative to physical distance, polymorphism is almost nonexistent. For example, in a gene in the base of the X chromosome, no polymorphisms were detected in 130 individuals from one species and one polymorphism in 109 individuals from another species. One possible explanation was a low mutation rate. However, comparison between species shows that this gene evolves normally. Divergence between species is about 10 percent, typical of other genes. These alleles show very low polymorphism but only because variation is not accumulating over time. In genes at the other end of X chromosome in *Drosophila*, chromosome recombination also is essentially zero and a series of genes

were selected in that region to more thoroughly investigate this issue. Non-coding regions of two genes were surveyed for variation using a stratified sampling (knowing the sequence, designing primers and then using SSCP). Very few polymorphisms were found. Even in these regions of low recombination, numerous cases show abundant evidence of recombination in the history. Still, for genes along the tip of the X chromosome, the level of variation goes steadily down. Recent data on human variation shows a significant relationship between the level of polymorphism and the level of recombination for physical length. Based on such studies, recombination rate may be important in understanding the rate of polymorphism.

It would appear that at some time in the past, some gene near the tip of the X chromosome was favorably selected perhaps because of a major evolutionary event and a particular allele swept through the population and went to fixation. Perhaps only one or two haplotypes survived or, even more recently, a single haplotype went to fixation.

One of the consequences is that there is less variation. Another consequence is that each mutation will be unique in a sample. Evidence has been found for that kind of distribution, with an excess of unique mutations in those samples.

These considerations indicate the importance of where variations are in the genome and in the gene. Candidate genes likely to interact with the environment will likely show a much more diverse distribution across environments and histories of humans than the average gene, a point worthy of consideration when selecting methods for sampling and analysis. However one identifies polymorphisms that will be used for typing, disease studies or anthropology, it is important that the ascertainment of these polymorphisms be well defined in terms of which population was surveyed to identify the polymorphism and the frequency of that polymorphism in populations of interest. Also important are the linkage disequilibrium relationships among polymorphisms. Numerous researchers who want to find polymorphisms conduct association studies but the problem for several genes is that a great deal of recombination in the history of the samples will reduce the association of a random polymorphism with a phenotypic effect.

How should the density of markers be decided? Regarding the distribution of relationships among different polymorphisms, at what distance genetically does the correlation drop off among sites? Theoretically it is expected to drop off precipitously and then stay flat. In the early 80's, the *white Drosophila* locus was surveyed densely to associate phenotypic expression with polymorphisms. Unfortunately, the linkage disequilibrium dropped off precipitously after two or three hundred base pairs, making sites separated by a few kb unlikely to be correlated and creating a very deep statistical challenge.

In flies, as in humans, the nervous system is patterned differently, and it has a certain heritability. Ten or 15 loci are critically important in determining how many and where the sense organs appear. On the order of 50 alleles for each locus were sampled and analyzed genetically to see their contribution to bristle patterning. The *delta* locus, a linchpin in cell-cell signaling, was an obvious candidate, with 57kb of known regulatory

and functional regions, including some coding region. Four different genetic backgrounds and 55 alleles were examined, using genetic crosses to identify the function of *delta* in different contexts. There was a serious problem with recombination. A similar obstacle faces human genome endeavors if, in fact, *delta* is representative of the kinds of statistical machinations required. Polymorphisms were found in the *delta* region that occurred at least three times or more in the sample. There was not much linkage disequilibrium in well over 1,300 tests. However, the few cases in which one found linkage disequilibrium seemed to be clustered among tightly linked sites less than a kilobase apart. Could one find an association between these markers and the phenotype? Half the time one might detect an association, by chance.

A practical approach is to compare DNA sequence haplotype data with phenotype data to search for an association. When a significant site is found, the question remains whether it is really significant. One approach is to permute all the haplotypes with respect to the phenotypes. The ability to make any particular statistical inferences is lost; nevertheless, an empirical distribution can be generated, and it is possible to determine whether a site is statistically significant or could happen by chance. Most importantly, this approach preserves whatever structure there might be in the data. An analysis of one bristle character in *Drosophila* looked at the scores for sites by the F test. One site in the second intron was quite significant. Following statistical regression to remove the effect of this one site from the phenotypic data showed another site that justified performing a new permutation distribution for significance. This revealed a site in the last intron with a very statistically detectable effect that is limited to only one sex, although it is a problematic variant that is undetectable in an individual fly. However, examining a large number of flies in a very specific situation, it is possible to detect about a 5 percent difference in the character.

In reviewing the approach described here, problems that emerge involve assembling the DNA sequences, detecting the polymorphisms and developing an inexpensive technology to type many individuals and to study populations. A major problem is the statistical one of whether population-based association studies can detect phenotypic effects versus more direct approaches to functional variation (e.g. transgenic mice). Whether these kinds of statistical techniques can be scaled up to the genomic scale is a very serious question. Very few investigators have the interest or training to perform this kind of analysis and the limiting reagent for genomic polymorphism and association studies will be trained researchers. Questions of the quality of the ascertainment of the sample, how to include the ascertainment in the analysis and then how to infer statistical significance, are extremely daunting, even in a laboratory animal under controlled conditions. In a human epidemiological setting, these problems seem even more daunting.

## Session III: Technologies

### Introduction

Sequencing technology presents three main targets. The first target is genes whose sequences are not known, which is properly the focus of the Human Genome Project. The Human Genome Project is enlisting techniques and technologies that, though powerful, are conventional. The sequencing is the "directed shotgun" approach, directed sequencing of clones. The project employs a variety of robots and other instruments. Up to now, the sequencing groups are accurate to about one error in  $10^4$  base pairs.

The second target, resequencing, involves detecting polymorphisms. This could be carried out by the conventional approach described above or by non-conventional approaches, such as chips, which will be covered more thoroughly in this session.

The third and final target will be typing those identified polymorphisms cheaply and quickly, which can be accomplished by sequencing, by chips or by other techniques. One of the other techniques under development by Glen Evans (University of Texas Southwestern Medical Center) and Kenneth Buetow (Fox Chase Cancer Center) points toward a method for typing polymorphisms with the potential to type tens of thousands of individuals.

Information on a group of possible therapeutically useful genes is already available—a thousand to tens of thousands of genes. A popular notion is to array these in a chip format. Evans and colleagues decided to use a not-yet-commercially-available technology to type genes implicated in environmental exposures. That technology is a DNA chip that fits easily on the fingertip. The chip, manufactured by Nanogen, offers a small array designed specifically for genetic diagnosis, not for resequencing. This chip has 25 test sites 20 microns wide, confined to a 1 millimeter sample area. Each test site is linked to an electrode imbedded in silicon. If DNA is present it concentrates on a probe at the electrode site, and is typed according to the polarity of the voltage applied. The chip fits into a machine, a chip-reader, that carries out an entire analysis in seconds.

The group set out to design a chip that would contain polymorphisms like NAT1 and NAT2 discussed earlier in the symposium, for their potential utility in environmental testing. Representative data for another allele of interest in environmental exposures, EPHX1, was presented to show how a potential 12-probe chip worked. EPHX1 is involved in detoxification of polycyclic hydrocarbons and in the metabolism and activation of dilantin. Two common alleles are known, one of which encodes an altered amino acid sequence that is associated with lower activity in the enzyme. It is suspected that this lower enzyme activity is associated either with effects of dilantin during pregnancy or with genetic susceptibility to an established environmental carcinogen. Thus, this would be a useful polymorphism to type in a large population.

The group generated polymorphic probes; oligonucleotides aligned in rows according to type. In the presence of the homozygote for A1, for example, it would have only a positive signal at a specific site; a homozygote for A2, only a positive at another site; a heterozygote, both positive; and if neither site hybridized, it would be indeterminate, meaning the test may have failed. The group targeted early 1998 to complete an entire chip, with 12 probes in operation at once.

## **DNA Sequencing**

Recent efforts to generate a map of the human genome have focused primarily on single nucleotide polymorphisms, SNPs. SNPs are the most frequent sequence variations and, thanks to digital technology, their detection offers the possibility of automation. How frequent are SNPs in the human genome? To answer this, a team led by David Wang (MIT, Whitehead Institute, Center for Genome Research) in collaboration with Affymetrix, a chip manufacturer, sampled 1,200 sites along a 300 kb segment of resequenced DNA. The SNP frequency was 1 for every 1,066 base pairs. More recently, the team resequenced 150kb from four ethnic groups of eight individuals each--Caucasians, Asians, Africans and Amerindians. In these groups they found twice as many SNPs, raising the question of how many SNPs will adequately cover the human genome. A series of additional studies at The Whitehead Institute suggested that a genome scanner of 1,000 SNPs and a genetic map of 2,000 SNPs would be sufficient.

Many techniques have been devised for detecting sequence variations—from single-strand conformation polymorphism and chemical mismatch cleavage to enzyme mismatch scanning and minisequencing, the latter favored by the Whitehead Institute's Center for Genome Research for its sensitivity and chip compatibility. The chips can count between 64,000 and 400,000 distinct nucleotide probes. One in four probes was a perfect match to the target sequence; the other three had mismatches. After hybridization, the matching probe at each position should have stronger intensities, the mismatching probes weaker intensities. If the target sequence continued to mismatch, a different pattern would result because the sequence would no longer hybridize strongly to the probes—there would be a strong signal at the polymorphic base, a weaker signal surrounding it. The design to genotype SNPs is the same, except that each allele is treated as an independent sequence.

Screening a genome for a large number of SNPs with ABI sequencing can be tedious and expensive, so the Whitehead group sought to identify SNPs directly on the chips by collecting scores of sequence-tagged sites, STSs, with short sequences across the genome. To find SNPs from eight people, they amplified STSs, then compared hybridizing patterns and designed a series of resequencing chips. 3.4 megabase of sequence, corresponding to 27 megabases of resequencing, was covered by 185-oligonucleotide chip designs. So far, the group has analyzed 1.8 megabases in each of the eight individuals and has found 1,659 candidate SNPs and a polymorphism rate similar to that found using ABI sequencing.

To scale up the matching system to 2,000 SNPs on a single chip, the group designed for each SNP locus a short PCR assay closely flanking the polymorphic base. To learn which loci were successfully amplified, they put the loci through a signal test and a sequencing test. They were able to amplify 20 to 200 loci at a time. Against PCR, 280 loci in an amplified pool of 553 passed the signal and resequencing tests.

Loci were also put through a cluster test involving 39 individuals to see whether the signals fell into two to three groups that would correspond to possible genotypes. Seventy-five percent of loci passed. Of those defined genotype clusters, further tests confirmed the system's accuracy and reproducibility—there was 99.9 percent agreement with genotypes independently determined by ABI sequencing.

The results show that a chip-based genotyping system can be accurate, efficient and reproducible. Such a system is critical for several applications, including large-scale family-based linking studies and large-scale population-based association studies. The work has yielded about 2,000 SNP markers across the human genome. The goal now is to construct a map of up to 3,000 SNPs.

### **Chip Technology**

David Cox and his colleagues at the Stanford Genome Center have produced 15,000 markers — sequenced sites (STSs) placed on a high-resolution radiation hybrid map. The markers can be placed on a chip and scores of SNPs can be generated. A large fraction of them, over 5,000, are genes; therefore, STSs that are close to genes can be useful as reagents since the STSs are already available and the sequence of the intervening material is available, not just at Stanford University and Whitehead Institute but at many other institutions, and they are in the public domain.

On an Affymetrix chip or another like it, it is possible to look at individual PCR reactions for candidate polymorphisms. Not everything on the chip turns out to be a polymorphism, and polymorphisms are probably being missed. But the Stanford group has found, comparing any two chromosomes, about one polymorphism in 2,000 base pairs.

What are the patterns of variation of these polymorphisms? In data limited to seven chromosomes across 30 kb, almost no disequilibrium was present. The variation pattern was of a nonlinear distribution, with no variation in the mid-section of the 30 kb. Disequilibrium doesn't have a predictable pattern. Haplotypes of those seven chromosomes showed that humans, being diploid, have an individual binary possibility for SNPs. But how would someone determine whether an AC and E were on one chromosome and a BD and F were on another? This is a problem with important implications for the Environmental Genome Project. CIPHER haplotype orders will be critical; samples must be collected in a way that allows this. Random individuals aren't sufficient; trios, parents and a child, are necessary in order to reconstruct haplotypes. Computer programs can reconstruct some haplotypes but are useless for rare alleles.

For the purpose of the Environmental Genome Project, it will not be possible to look for all of the genetic variation. Because of cost, the picture will be incomplete. On a single chromosome, one could find every variant. But population geneticists will want to look at many chromosomes and search for variants. If that number is 10 chromosomes, there will be about three times as many variants found. At 100 chromosomes, there will be five times as many variants found. With a billion chromosomes, there will be more variants discovered. The progression is nonlinear because many of these variations are rare in the population. Some researchers have advocated looking for rare ones; some have advocated looking only for the common ones. Some have pushed for looking at variation in different ethnic groups. Others want to focus on people who are socially disadvantaged.

Because all of the variation cannot be examined, project participants will have to accept that they have to come up with associations and other biological interpretations with only part of the information. They will have to estimate—for instance, one base change for 2,000 base pairs, or what one would expect with an infinite allele model for a hundred chromosomes. Using this model, at 5 kb there would be 13 base changes on a hundred chromosomes. At 10 kb, 26; 20 kb, 52. If there were no recombination, how many haplotypes or alleles would be seen? This will be a difficult question to answer, and the answers will be different in different regions of the genome. For example, in 100 kb at 1 percent frequency base changes and higher, only two-thirds of all the variation would be accounted for. The take-home lesson: whether by DNA sequencing, or even by another more perfect technology, only a portion of the variation will be found.

This project, then, should concentrate on the DNA base changes that are common across different members of the population. An example from LPL discussed earlier in the symposium and another from the Stanford study of chromosome 21 indicate that some variants are present in one ethnic group but not in another. Yet a significant fraction—between 30 and 50 percent of all variants—are present in every ethnic group.

Because the purpose of the Environmental Genome Project is to improve the health of people in the United States, finding variants common to everybody in the country would help everyone. Concerns about including one group or excluding another are moot. Because only a fraction of the polymorphisms can be found, the fraction present in everybody—“everybody” defined as a sample of the U.S. population—would be a representative U.S. sample. Might this apply even to the rest of the world? Probably. Though Americans are not identical socially to everybody else in the world, they are genetically representative of many of the world’s peoples.

Should the project consider only genes that are already sequenced—that is, should the project focus on resequencing? There are a variety of screening techniques that start with small segments of DNA, either PCR-amplified or cloned, and allow identification of those segments that contain single base pair changes. For example, it would be possible to clone a gene and compare two genes divided into small pieces of 200 base pairs each. Based on the premise of only one base change every 2,000 base pairs, the two genes

divided would yield a variant ratio of 1 in 10. Identifying and sequencing the variant pieces would therefore require sequencing only 10 percent of all the material. These technologies are only screens, it should be noted. They will not find everything, but they do provide a way of using partly sequenced DNA to identify variants that can be confirmed by sequencing. It would be an inexpensive approach and a way to look at unsequenced genes of biological interest.

One approach Stanford has experimented with is a bacterial color assay. The assay uses mismatched-repair insertions of three base pairs, *in vivo*, to identify single base pair changes in DNA heteroduplexes. The sequence of the clone that is used is unknown. If there is no mismatch, the bacterial colony is blue. If that inserted sequence has a mismatch, the bacterial colony turns white. Thus, the color assay screens for the fragments in question, the ones that will be selected for sequencing to identify the polymorphism. The technique is not new—it was published in the December 1995 issue of *Genome Research*, to demonstrate a single-base-pair change in a mouse gene involved in neurological function. It was a single allele, and there were many candidate genes, but no one had identified the mutation. In this instance, the technique identified a single-base-pair change, from G to A at nucleotide position 953 in a potassium channel gene. In mice homozygous for the mutation, a candidate gene was cloned into the vector, as was wild-type mouse DNA. The wild-type alone showed 37 percent white colonies. The mutant mouse with this particular gene had 40 percent white colonies. But in the heteroduplex of the wild-type and the mutant when there was a mismatch, 86 percent of the colonies were white. Sequencing confirmed that this was the gene with the change. It turned out that because of the mutation, this channel no longer transported potassium, it transported sodium. In 1.4 kb of DNA, a simple screening test showed that this was the gene that should be sequenced from a whole list of candidate genes, and the same approach can be used for polymorphisms.

A final issue involves open disclosure and intellectual property. These techniques cannot be used in humans if their inventors keep them secret. Nor can they be used if they are so encumbered by licensing agreements that no one can gain access to them in a timely fashion. Something akin to this is happening now with SNPs. In fact, many SNPs that have been associated with biological functions have been patented, so there is no question about whether SNPs can be patented. The question is whether they should be. This is a critical issue that needs careful attention now, before thousands of polymorphisms are licensed in a way that seals them off from the world.

## **Session IV: Population-Based Epidemiological Studies**

### **Introduction**

To accomplish its goals, the Environmental Genome Project requires a strong epidemiological component. The point of identifying polymorphisms in potentially

important environmental-response genes is to improve the ability of epidemiology studies to identify health effects due to environmental exposures. There is a dearth of functionally important and relevant polymorphisms available to be incorporated into epidemiologic studies. The Environmental Genome Project will fill this gap.

Epidemiologists have a number of important roles in this project, including nominating genes for study by noting differences in disease rates across populations and by identifying potentially interesting candidates. A second role for epidemiologists in the project, which may not be obvious to them, is in clarifying the functional significance of some of the polymorphisms identified *in vivo*. The third and obvious role is to use polymorphism data to investigate their role in environmental disease. An important issue is how to optimize collaboration between disciplines; not only molecular biologists and epidemiologists, but quantitative statisticians and classical geneticists as well.

Other issues are the potential of chip technologies to change how scientists in various disciplines interact with one another. Assuming that chip technologies do become cheap and amenable to large-scale studies, it would be technically and economically feasible to treat them just like any genetic marker and look for associations. A second, more intellectual approach would be to wait until the functional significance is known for all these polymorphisms and an extremely strong rationale has been established for an environmental disease relationship. But functional data are much more difficult to generate and those will be slower to accumulate, so there may be a hazard to being so careful and waiting.

There are a number of possible study designs, which raises the question whether traditional cohort and population-based case control studies are adequate to determine the role of various polymorphisms and the need to consider alternative designs. Studies of gene-interactions require large sample sizes, even when looking at just one gene and one environmental exposure but especially for multiple genes and multiple exposures. A simple example: detoxification and activation steps, which involve at least two genes and potentially many more. To get large sample sizes with sufficient power to determine gene-environment interactions, one could perform a single giant study, say 14,000 cases of ovarian cancer and an equal number of controls. Another approach would be to encourage investigators with samples already in the freezer to cooperate or, alternatively, combine individually published data using meta-analysis techniques. Another approach would require or request that investigators contribute samples to a central repository. Discussions should address the feasibility of these various options. Reams of anticipated data will require ever more personnel. Issues of interpretation are not simple and potential issues of publication bias arise regarding chance positives.

Issues for discussion include encouraging interdisciplinary collaboration, appropriate approaches to investigate newly discovered polymorphisms, empirical versus a more careful mechanistic approach, study designs and power and how to make best use of data.

## Opportunities

Misleading attacks on epidemiology and its methods have emerged over the past 10 years. The nub of the argument is that epidemiology is an observational science; only experiment matters; therefore, epidemiology is not of very much use. Whether epidemiology is an observational science is not particularly relevant. The key issue is measurement and its improvement through proper design.

That epidemiology is largely an observational science is true. But so, too, are cosmology, astronomy, evolutionary biology and human molecular genetics. Most scientists think they are performing experiments but generally they are making observations, albeit using some sophisticated techniques. Unlike clinical medicine, biochemistry or molecular biology, epidemiology involves the study of populations or groups, and therein lies both its uniqueness and some of its interesting characteristics. There are barriers to communication between epidemiologists and other scientists, particularly those in molecular genetics and molecular biology. They have problems not knowing each other's vocabulary, sometimes not understanding the underlying concepts, problems with misusing the vocabulary and using different words for the same concept.

Developing studies with more homogeneous exposure allows the identification of interactions and therefore helps define the mechanisms of susceptibility and exposure. For example, data from a study associating physical activity, energy intake, body mass index, or BMI, and the risk of colon cancer turned up markers for exposures that are possibly genetically determined in 2,000 cases and 2,000 controls. Self-reported measures of physical activity of “high,” “intermediate” and “low” were combined with measures of energy intake. The referent group is people who have low energy intake, high physical activity during their lifetimes and body mass in the lowest third of the population. As body mass increases in this population at the lowest energy intake and highest activity, the data indicate that body mass has no effect on the risk of colon cancer. However, in the high energy intake group, there is perhaps a stepwise increase in risk, although again no association with BMI. People with the highest energy intake, lowest physical activity and highest body mass—the ultimate couch potatoes—show an approximate four-fold increase in risk of colon cancer. BMI may be determined in part by the relationship between energy intake and energy output, but risk is determined by all three variables. When this large population is stratified on a biologically meaningful variable like gender, rather than merely a four-fold increase in risk, the men demonstrate about a seven-fold increase in risk. Intriguingly, there is almost no association for women.

The capacity to stratify raises interesting questions and may implicate some genes. Studies like these have significant power when a population is sufficiently large. Using the same analysis for a polyp study, exactly the same pattern emerges in which men have a 3.5 to four-fold increased risk when they have high BMI, high energy intake and low physical activity. Again, in women there is essentially no association for the same set of variables. Assume one approaches this stratification issue in genetic terms. There are no

genes for physical activity, as far as we know. Are there genes for energy intake? That is problematic. Whether or not genes influence body mass, one still faces the problem of defining why men and women are different.

A nearly complete study on NAT2 and polyps indicates an imputed phenotype based on genotyping the study population. The genotypes have been stratified into slow, medium and fast NAT2. Analysis of smoking is incomplete although it is clearly associated with an increased risk of polyps and risk is highest among heavy smokers with fast NAT2, with approximately a nine-fold increase in risk compared with the slow NAT2 individuals who have never smoked. Here, too, one benefits from thinking about the interactions but fairly large numbers are required to make meaningful estimations.

Numbers aside, whom do we study? An interesting family is good for gene hunting but not very useful for population problem studies, for studying the full spectrum of gene-disease associations; for understanding gene-gene interactions; or for fully understanding gene-environment interactions. Large numbers of self-selected individuals are not useful. Perceived cancer risk in family history will ensure the genotypes being measured will be biased. What about population-based family study designs that have control of environmental data and the genotypes or segregation analysis leading to linkage? What are the best prospects for success? The advantages of population-based designs, out-bred populations, gene-carriers with no disease and interesting families with low penetrance, are absent if a design selects only high-risk families.

Regarding defining genotypes, although chip technology will simplify studying large numbers, there are some caveats. Our methods for defining phenotype have been inherited from hundreds of years of medicine--diseases defined in highly heterogenous ways; anatomy, microbiology, histopathology and so on. Chips will provide a large increase in data on genotype or gene expression. But how much error is acceptable? If genotype is defined by a single gene, associations of the disease will be no more problematic than current epidemiologic designs. In fact, a 90 percent determination of a specific genotype will be much better than self-reported data and, occasionally, better even than biochemical measures. However, if the genotype is defined by 10 genes, a 10 percent error results in correctly genotyping just 35 percent of the population. When defined by 100 genes, accurate genotyping occurs for only three in 10,000 people. Assuming error improves to only 1 percent, for 10 genes accuracy goes to 90 percent but for 100 genes, it is 37 percent. For 6,000 genes, which happens to be that of the yeast genome, accuracy drops to .2 percent. Therefore, we need to address how we define genotype in populations and with what degree of accuracy.

The body is made up of from  $1 \times 10^{12}$  to  $1 \times 10^{15}$  cells. Most cancers are clonal. Fewer than one-third of the population gets cancer, meaning one cell goes bad successfully in a total of  $3 \times 10^{14}$  genes every 75 years: a rare event. A major difference between cell biology, animal carcinogenesis or therapy and epidemiology is the frequency of events. In cell cultures, millions of abnormal cells appear. In animal experiments, between 30

and 100 percent of the animals get the disease under study. In a cancer treatment unit, everyone will have cancer. But in a population of 100,000 there will be fewer than 100 cases in a year for any specific cancer. To make the point in a different way: If one enrolls 4,000 men, age 50-69 in a cohort study, after five years, 64 lung cancers will occur in that population, about 1.6 percent. What if they are all heavy, 20-plus pack-years, smokers? What proportion of this population will get lung cancer? The answer is only a few more, about 2.5 percent. These are rare events even in people at very high risk of disease.

Given the rarity of events, there are two strategies. One can increase the numerator and ignore the denominator or increase the denominator to insure an increase in the numerator. Here are some biologic disciplines that increase the numerator: molecular biology, cell biology, animal carcinogenesis, family studies and most clinical studies. Epidemiology is the only discipline that uses the approach of arriving at an estimate in the population by increasing the denominator rather than the numerator.

Returning to the notion of improved measurement, one should look in the right place. Comparing astronomy and epidemiology, astronomy began long ago when men and women peered up into the sky and made observations with the naked-eye. Observation improved once the telescope had been invented, but there were false leads—for example, the canals on Mars. Observation continues with improved instruments. Similarly, measurement improves as we go from self-reported epidemiology to self-report plus biologic measures; from stratification on disease subtype from histopathology to somatic genetics; from family history alone to family history plus germ line genetics.

The course of progress towards understanding the involvement of human papilloma virus (HPV) in the origin of cervical cancer is a good example. In 1973, Rotkin observed that multiple partners are associated with an increased risk of cervical cancer; in the middle 1970s HPV was identified as a possible causal agent. Then in the 1980s, a series of cases showed HPV in cervical biopsies and further supported its appearance as a causal agent in the cancer. An understanding developed of the biology HPV16 and its ability to transform cells, and in the late 1980s, methods were available to avoid the problem of misclassification while measuring HPV in the field. By the 1990s, reliable field methods became available for HPV. In 1973, multiple partners increased risk by about two-fold for cervical cancer but by the mid-'90s, HPV was associated with between a 16- and 60-fold increase in risk, and when defined specifically in terms of HPV16, between a 30- and a 300-fold increase was found. So the precision of estimates of relative risk have improved spectacularly over just twenty years by virtue of improvement in biological measurement. Thus, biological insights frame the questions but epidemiological insights and methods are necessary to answer those questions. In short, think biologically but act epidemiologically.

## Optimizing Study Designs and Power

Historically, geneticists have had limited interest in the environment and have pursued approaches such as genome-wide scans and positional cloning to attempt to find disease genes. Epidemiologists, on the other hand, attempt to understand the function of a gene and characterize it in terms of penetrance and gene-environment interactions. An epidemiologist normally begins his or her career with an interest in a particular disease, trying to discover exposures related to that disease, and only then starts consulting with geneticist colleagues. That might be characterized genetically as a looking-under-the-lamp post phenomenon because, according to intelligent strategy, one waits for data about functional significance. However, one can argue that there must be a better way to find genes that interact with environmental agents.

A case can be made for using family studies rather than population controls to quantify genetic effects and gene-environment interactions. In one example of this, the amount of allele sharing within affected sib-pairs has been calculated for a case in which there is a gene-environment interaction. Using a modest two-fold relative risk interaction, more allele sharing is expected within affected pairs, both of whom are exposed, than in those in which neither is exposed or discordant pairs. If the environment is ignored, the epidemiologist would never look for this gene and would miss this gene-environment interaction. Tests can be constructed to take into account environmental co-variants; examination of just the exposed pairs is the most powerful approach.

Three basic types of sib-pair comparisons include: comparing the cases with their unaffected siblings for evidence of association; linkage analysis on pairs of affected cases; or, third, comparing cases against their parents, which can be informative about both linkage and association. All three are informative about gene/environment interactions and they are most powerful in combination with each other. Despite the effectiveness of sib-pair linkage for finding major susceptibility genes, the sample science requirements for finding common metabolic genes are astronomical. Nevertheless, such genes can be identified by association studies.

There are three basic association study designs. Case-control designs involving population controls; variants of family-based case control studies; and something only relevant for gene-environment interactions, a design that uses only cases. The latter will say nothing about the main effect of either the gene or the environment but does allow inference on the gene-environment interaction, assuming independent distribution in the source population. For example, a recently published study reported on the association between BRCA genes 1 and 2 and oral contraceptive use. The case in this design is a gene carrier and the control is a non-affected non-carrier. One sees a very strong association, albeit based on a small sample. The obvious interpretation would be an interaction between oral contraceptives and breast cancer genes. But we have to question the independence assumption. The genotype and oral contraceptive use could be associated, probably not directly, in that the breast cancer gene causes one to use oral

contraceptives, but indirectly. The BRCA1 gene certainly causes breast cancer and family history could modify whether a woman decides to take oral contraceptives. Case control data shows that such associations appear to be weak in the source population.

Clearly, we would like to interpret an association as a causal effect of the gene on disease risk. But alternatives should be considered. The first, a non-causal but interesting association, would be due to the actual causal gene being in the same region and in linkage disequilibrium, which might help narrow the search for that gene. Frequent associations are found, but, because association is with a different allele in different populations, results are inconsistent. The literature is littered with examples of disequilibrium with unlinked genes due to the phenomenon of population stratification. Convenient samples, such as clinic patients down the hall, or estimates of allele frequency derived from a resource such as the one that the Environmental Genome Project is proposing to establish, can lead us astray.

To the epidemiologist, ethnic stratification is doing nothing more than confounding another risk factor called ethnic origin, which is related to disease allele frequency and an independent risk factor for the disease. A prime example of such associations relates to the dopamine gene and alcoholism. Initial reports showed 77 percent of alcoholics carried the A1 allele at the D2 receptor locus compared to 28 percent of controls in a sample not well-matched on ethnic ancestry, although it did have roughly similar black-white distribution. This finding led to a large number of follow-up studies, all of them conflicting. The gene frequency is highly variable, often with bigger differences between studies than between cases and controls. More importantly, studies looking more rigorously within homogeneous populations have failed to find such an association. Furthermore, linkage studies and transmission tests also have been negative and, to date, no studies have been done with family controls. A second example involved non-insulin dependent diabetes (IDDM) in a Native American community and an apparent strong relationship between a particular hemoglobin haplotype and IDDM risk. But that association is not seen within full-blooded Native Americans. It turns out that this particular haplotype is nothing more than a marker for Caucasian heritage, which is the real risk factor.

The dangers of population controls are magnified by the issue of studying relatively rare genes and even genes of 1 percent population prevalence. In an ordinary population-based case control study, the small yield of gene carriers in the controls results in an unstable estimate of relative genetic risk. This argues in favor of a multistage design—for example, a matched-case control study matched on family history, genotyping only the family history of positive pairs and sampling family history negative pairs. This can be a very efficient design. Alternatively, one could start with an existing matched-case control series and genotype only the family-history mismatched pairs, a design called counter-matching. Though counterintuitive, studying mismatched pairs is both an unbiased and highly efficient design.

Family studies offer significant advantages. Though using data from families, the aim is to estimate population parameters. Family studies avoid concerns about bias due to population stratification. Disadvantages to family-based studies include a smaller pool of potential controls and some loss of efficiency due to overmatching. Nevertheless, a highly efficient design may be obtained by restricting to multiple-case families.

From the parents of cases one can construct a set of pseudo-sibs—including all the possible genotypes of the offspring of these parents, even those not included in the actual sib-pairs. The efficiency of these designs is improved by increasing the proportion of susceptibles in the sample but using care. The basic rule of thumb is that any restriction applied to the cases has to apply to the controls. If the case is required to have an affected offspring or affected mother, so too should the control. Otherwise, the approach would lead to biased estimates of the relative risk. However, the bias can be more subtle. Suppose the case's mother is a relative of the control but the control's mother is not a relative of the case, which would create bias toward the null and vice versa.

There is a wide range of valid studies. For parents of siblings, if the case has an affected mother then, by definition, the control does; similarly if she has an affected sister. If the sister is available for genotyping, a more efficient design results because two cases are compared to one control rather than just one. If the relatives are cousins and they share a common grandparent, the design is valid.

Numerical comparisons show the relative efficiencies of designs under a variety of genetic models. Though family-based controls are less efficient than population controls because they are overmatched on genotype, they do have protection against population stratification. Given a choice between bias and efficiency, most would prefer an unbiased design. Cousin controls are more efficient because they are less closely matched. Pseudo-sib controls are about the same efficiency as population controls but considerably more efficient in the case of a major recessive. The requirement that a sib-pair have an affected sibling or parent produces a design considerably more efficient than population controls.

These calculations also can be performed for gene-environment interactions. Sibling controls, though they are overmatched both on genotype and on environment, nevertheless prove extremely efficient for gene-environment interactions. This is counterintuitive until one reflects further. The most informative types of sib-pairs possible fall into three groups: when both are carriers but discordant for exposure; those concordant for exposure but genotype-discordant; and those discordant for both. A partial matching on genotype ensures that sib-pairs will tend to have a much higher yield of carrier-carrier pairs; more efficient than population controls in which the distribution of genotypes should be independent. The most informative group is carrier-carrier sibs who are exposure discordant.

Intelligent use should be made of large, heavily informative families: the Breast Cancer Linkage Consortium, which supplied data needed to find the breast cancer gene; the various NIEHS efforts; the familial cancer family registries; and the cancer gene

networks. Families in the registries will be highly informative in terms of other family members' phenotypes. The group of Duncan Thomas at the University of Southern California has been exploring partial segregation analysis on only those genotypes compatible with those of observed, typed individuals, resulting in dramatic improvements in efficiency. This requires a population-based series of families.

A major challenge remaining is the highly polymorphic gene. Numerous genes proposed to be studied by the Environmental Genome Project include many different known alleles. With small data sets and multiple comparisons, it becomes difficult to decide whether a variant is a neutral polymorphism or has some impact on penetrance. The USC group has been considering the use of hierarchical base models in this context.

## **Session V: Functional Analysis of Polymorphisms**

### **Introduction**

An alternate name for this session might have been "How do we determine the functional significance of a polymorphism?" Functional characterization of a polymorphism is an extremely multifaceted and challenging problem. In metabolism genes, there are many known single-nucleotide polymorphisms that occur in coding regions and exons. Yet, there are other types of variations as well, such as frameshifts, deletions and amplifications, and so forth. These examples show that functional polymorphisms will not be limited to coding regions or to single nucleotide changes.

Generally speaking, how might one go about working up the functional significance of a polymorphism? A polymorphism can cause the activity of the gene product to either increase or decrease or be unchanged. Additionally, the polymorphism could lead to deregulation of the expression of a gene product. Another well-documented mechanism is "dominant inhibition" – the manufacture of an altered protein or nucleic acid product that interferes with a pathway, leading to a deficiency in the pathway. Altered enzyme specificity also must be considered. For example, if RNA polymerase is altered in its specificity so as to be error prone, one might expect a major change in cell phenotype. More directly, the analysis of tissues and cells harboring a polymorphism might seem to be a straightforward way to start.

Quantity and stability of a protein can be measured, as well as the *in vivo* activity in the cell. For example, a metabolism enzyme could be tracked systematically by this technique. Finally, a protein's biochemical activity in an *in vitro* system can be studied, where an enzyme might function the same way it functions in the intact cell. There are abundant examples of this: protein synthesis, RNA synthesis, DNA replication, drug metabolism. The list goes on. In this approach, it is implicit that we have enough information about structure-function relationships of the wild-type gene product so that reasonable guesses can be made as to what the polymorphism might mean in terms of a functional alteration.

Another topic has to do with knowing enough about the gene and the gene product to design straightforward assays. Study of protein turnover or enzymatic specificity using available antibodies, cDNAs, and expression vectors could be the most facile way to look at enzyme function. However, because a polymorphic enzyme's turnover number as a purified protein in solution is understood and does not show any change relative to the wild-type enzyme, this does not necessarily mean it will function properly *in vivo*. The complex macro-molecular assemblies that occur in the cell will need to be reflected in the *in vitro* assays as much as possible.

In keeping with the concept of macro-molecular assemblies, one must consider protein-protein interactions. Many proteins obviously act in partnership with other proteins, sometimes with more than one. Therefore, it is important to include assays that allow analysis of protein-protein interactions.

Ultimately, use of genetic systems such as yeasts, mammalian cell lines and animal models, are very important tools to consider, because in most cases we will not know enough about the role(s) of a protein in the cell to understand the functional significance of a polymorphism through biochemical analyses alone.

For some proteins that will eventually appear on the Environmental Genome Project list, structural determination for a polymorphic gene product is not beyond reason. If proteins that have already been analyzed extensively at the level of three-dimensional structure are given priority for the Project, our ability to focus the structure-function work-up will be greatly facilitated. Theoretical analysis of the polymorphism can also be undertaken, and much more research and analysis is required in this area.

What about mRNA expression? Obviously, it will be important to characterize polymorphisms in promoters or other regulatory regions in model systems. Theoretical/computational analysis is an approach that merits consideration here also. Sequence alteration of many promoter elements renders them inactive in promoting an RNA synthesis event. The same may hold true for other types of sequence-directed binding events or enzymatic events involving, for example, RNA- and DNA- binding proteins.

Let us also consider the topic of cellular response which is an area of great potential. If a polymorphism changes cellular metabolism, triggering a stress or alteration response, mRNA pattern recognition of such responses may offer a shortcut to understanding a functional polymorphism and predicting what the biochemical mechanism might be.

Altered metabolism is an important approach to consider as well. With many gene products, overall metabolism in the cell could be ascertained simply by making a cell extract, quantifying metabolites or looking at metabolic pathway products.

Finally, the role of population-based studies in assigning function must be addressed. It may be wise to consider population-based studies as an integral part of the systematic work-up of a polymorphism.

The first speaker in this session covers this entire area of functional analysis in detail. The second speaker focuses on the topic of the quantity and stability of cellular mRNA, a potential exposure marker that can be measured using exciting new technologies. In most cases, the pattern of expression of an mRNA can be measured routinely as a function of cell-cycle stage, for example. Yet, new technologies promise to provide greater capability to carry out mRNA analysis for many genes simultaneously.

### **Functional Analysis Models**

Functional analysis is extremely complicated and problematic. Looking for variations in enzymes and other gene products that carry associated disease risk begins with assaying function. The endpoints are clear in some cases but unclear in most. Of particular concern is whether these studies are relevant for target organs. If there is an environmental component, what is the effect at a disease's onset? This remains a problem because exposure may have changed during the course of the disease, and that information may not be available. And many environmental diseases of interest are due to very complex mixtures—in many cases, no one really knows what the major components are, but assays must be devised nonetheless, often not involving the mixtures themselves but individual components, so working out the etiology is uncertain. Because of this, the basis for effect may not be established going into a system, yet predictions will be based on a functional assay. Another problem concerns enzymes that metabolize xenobiotic chemicals. Often, one is dealing not so much with one enzyme but actually with a balance of different enzymes, some of which activate, some of which do other things.

Defining polymorphisms in the population is really the first step. There's no question that more polymorphisms will be found; the question will be how to determine which ones are important. There are two major ways to get at this question. One is to do associations with the epidemiology; the other is functional analysis. A combined approach is necessary.

While function cannot be ascertained by epidemiology alone, epidemiology is a necessary component in this process. An analogy is the process of drug development in the pharmaceutical industry. A Phase III clinical trial is necessary before a drug can be put on the market. On the other hand, the role of functional analysis in drug development, the *in vitro* screening, involves coming up with systems that predict which drugs are most likely to be efficacious in people and which drugs will get through the clinical trials and become useful new medications.

Getting down to basics, the significance of a polymorphism depends on the phenotype. Under many conditions, a genetic change can lead to phenotypic changes. For instance, if a new codon gives an amino acid substitution that changes the property of the protein, and if this is observable, a phenotypic change occurs. Now, suppose one inserts a codon or modifies a residue to get an altered codon and no protein results. RNA transcription also can be decreased or, in some cases, increased. And there can be variants that cause improper or inefficient RNA maturation. But to find out, one must have an assay.

Without an assay, the situation is akin to the proverbial tree falling in the forest with no one to hear it.

What are the general strategies for assaying function? There are a couple of different possibilities. One is to do *in vivo* human assays with clear endpoints, which should be possible with some enzymes, particularly those involved with metabolism. For instance, with cytochrome p450 one can actually use drugs as markers to see if there are any real changes in people—information relevant to drugs. Can this be done for diseases with complex etiologies or unknown causes? The other possibility is to express the polymorphic allele in a heterologous system and then examine function. The idea here is to use a specific assay. For example, if one is looking at glutathione peroxidase, there is an established enzyme assay with high specificity. The question now becomes whether the results obtained in the model system has any relevance in a more complex system. In some cases, the pathology of the whole cells or organism will be of interest, so the main question will be which model system is appropriate to study the human disease.

Model systems for heterologous expression vary considerably, including microorganisms like bacteria and yeast, mammalian (including human) cells, and transgenic mice. Most cell-based systems of interest here utilize tumor cells, either with transient expression or stable transfection. There may not be a single system that works for everything. Bacteria won't be useful in most cases, but there may be a few instances involving well-characterized enzymes where they could be. A lot of time will be budgeted toward figuring out what system will be easy to use to express many constructs. This project is looking at approximately 200 genes. Being conservative, each of the genes might have 10 polymorphisms of interest—that entails expressing 2,000 genes. Not just cDNAs. Another problem is the endogenous background. For instance, if a transcription factor is expressed, what is the effect of the one normally in this particular cell line? The same question could apply to a metabolism gene. Must one knock out that endogenous gene first to study the heterologous gene? Another issue is the similarity to humans. Is genetic regulation going to translate from the model system? Some will argue that yeast isn't the appropriate model. Is a sick yeast representative of a sick person? Will proteins look the same in the model system and humans? What about interactions with other components. This gets complicated when things that can function as transcription factors are thrown into the equation.

Complexity, in fact, is a big issue. What are the active components in such complex mixtures as tobacco smoke and smog? Should systems be set up where tobacco smoke can be put into cells, or should the focus be on polycyclic hydrocarbons as indicative of what's happening with the total mixture? What about the age-old issue of dose, about what dose will accurately reflect human exposure? Will so-called intermediate endpoints, such as DNA adducts and mutations, necessarily be relevant to a particular cancer to be studied? Also, unexpected new functions might not be anticipated in standard assays.

With all these obstacles, it might seem as if there were little hope of understanding the influence of genetic polymorphism on disease. But there is a bright side. First, history can

be a good guide. Consider the field of human genetics in inherited disease—it has led to the understanding of many complex diseases, such as metabolic diseases associated with single mutations. And numerous animal models have demonstrated susceptibility to cancer, chemical toxicity and other diseases in which genetic differences, polymorphisms, do crucially interact with environmental agents. Understanding drug metabolism has led to therapeutic drugs. It is well established that there are wide variations in enzyme activities in humans and that there also can be dramatic *in vivo* differences in drug-induced toxicity. With metabolic enzymes, this can be pinned down, and perhaps these same enzymes can be useful in studies of environmentally-associated diseases.

The Environmental Genome Project could take two approaches toward functional analysis. The working group could set protocol for a common approach to everything, or it could encourage multiple approaches to individual genes. A universal approach to functional analysis would confer advantages. One system might offer utility for cross-comparison. For instance, the project might choose to express everything in yeast—this is just a hypothetical suggestion, not a recommendation. But in this hypothetical example, it might be possible to measure enzyme activity, characterize the yeast phenotype and quantify this information, then try to relate it to human epidemiology. It may be more realistic, however, to perform a battery of different approaches that are system-based. Consider, for instance, enzymes involved in the metabolism of xenobiotics—e.g. cytochrome p450 enzymes and glutathione transferase. These have been studied extensively. Once polymorphisms are known, it may be possible to correlate these polymorphisms with *in vitro* catalytic and tissue samples. Ultimately, it might be possible to express proteins and measure function in humans (not, however, in cases of environmental exposure, because carcinogens and toxins cannot be tested in people).

The scenario gets more complicated when DNA repair enzymes, polymerases and transcription factors enter the picture. Then there are tumor suppressors, cell-cycle checkpoints, signal transducers—functions whose role in environmentally-associated disease may be difficult to determine. And in how many model systems should these functions and interactions be examined? There's a tradeoff: the project can generate a lot of information about everything, or it can be streamlined to prioritize and carry out only the most important assays, and then attempt to correlate the laboratory results with supporting epidemiology. These are decisions that must be made, in order for this work to be accomplished.

### **Gene Expression Assays**

Tools being developed for the Human Genome Project to inexpensively analyze a large amount of genetic material may also be useful in the arena of studying gene-environment interactions. One such promising technology is the DNA micro-array, which Patrick Brown and his group at Stanford have used extensively. The approach may yield more useful information at a much lower cost than would resequencing 1,000 alleles of 200 different human genes. The DNA micro-array can display thousands of different DNA

sequences. Thus far, the Stanford group has assembled arrays that contain every known gene in yeast, a little over 6,000. For humans, they have arrays that contain more than 10,000 genes. In the next year, they hope to be able to look at 40,000 human genes at once.

The arrays are noteworthy for their simplicity of construction. Their use requires a robot, made with “off-the-shelf” parts at a cost of less than \$20,000, and a well-trained graduate student or post-doctoral fellow. The robot can print 100 yeast-gene arrays at a time, in a little less than two days. An array under development will be able to accomplish the same task in about an hour and a half. Each array contains 6,400 DNA elements, including gene sequences and controls, allowing the user to gather quantitative differential abundance data of specific nucleic acid sequences. The use for such a technology is open-ended; specific uses rely mainly on the investigator’s ability to figure out how the array might be used to derive information on functions.

An obvious example might be to ascertain which genes are differentially expressed in a tumor cell compared with its progenitor cell. For that matter, one could look at which genes are expressed in response to environmental stress by comparing cells before and after exposure to oxidative stress or, for example, to water from the Potomac.

To do an experiment, total mRNA is isolated from each of the two cell types. Differentially labeled cDNAs are prepared from each of the two mRNA populations, and the samples are mixed and hybridized to a micro-array. The labeled sequences hybridize to a cognate spot specifically, fractionating the mixture gene by gene. Because the relative amount of the labeled cDNA sequences is compared on the same slide, factors that would normally complicate measuring differential expression are eliminated. On a typical array, each fluorescent green or red spot represents a different gene. The relative amount of green versus red that hybridizes to a spot yields a highly accurate readout of the relative abundance of the mRNA in the two samples. Looking at the cancer cell example, a yellow spot is a gene expressed to an equivalent level in the two cell types, a green spot is a gene expressed at a higher level in the normal cells, and a red spot a gene expressed at a higher level in the tumor cells. Using image analysis programs, the pattern can also be quantified to aid further interpretation.

The Stanford group is pursuing several applications for micro-arrays that include mapping protein distribution, conducting large-scale genotyping, detecting amplification or deletion of genes and carrying out genetic screens—which could include identifying genes that interact with environmental agents.

In fact, gene expression patterns will point to pathways that will be important in toxicity or resistance to environmental agents. The micro-arrays can assist in screens for those genes that can influence susceptibility or resistance. The technology can be useful from the standpoint of looking for biomarkers in populations susceptible to certain environmental exposures. The micro-arrays may supply gene-expression fingerprints from tissue samples that will show defined patterns associated with exposure to

environmental agents. If the patterns can be recognized *in vitro*, they can be used to monitor patients for research purposes or, in the long run, to prevent exposure in the population.

An example from yeast, looking at more than 6,000 genes, shows the power and the utility of the micro-array in assessing what would be akin to an environmental catastrophe—a cell deprived of sugar. The Stanford group grew yeast in glucose until the yeast exhausted it and converted it to ethanol. The changes occurring in the cells were observed over time to determine a detailed profile. When the yeast ran out of food, growth stalled drastically. Afterward, in the micro-array, many of the 6,000 genes showed a significant alteration in expression, which was reflected in bright red and green spots. In fact, more than 25 percent of the yeast genome—16,000 genes—showed at least a two-fold change in their expression level, and 7 percent showed at least a four-fold change in expression.

With the same assay, the group was able to generate a detailed picture of what was happening with the transcription level of enzymes in metabolism genes and to look at expression under a variety of other conditions. For each condition, they could see a distinctive pattern that could provide a diagnostic tool for recognizing the state of the cell; one pattern, for example, would signify the loss of function of one gene and the activation of another. This creates the possibility that the patterns could be used diagnostically—by observing familiar patterns from a cell exposed to unknown agents, one could at least ascertain the state of the cell. This might turn out to be useful in the Environmental Genome Project. For example, there might be a genetic variant in a gene that is implicated in resistance to environmental hazards, or a specific pattern could emerge from a change in environmental exposure.

The Stanford group, in fact, has assembled micro-arrays to illustrate that there is no difference in the method whether one is analyzing yeast or human gene expression. In one study, for example, they looked at the changes in expression in growth-arrested fibroblasts after exposure to serum. After 15 minutes, they could see that many genes were highly induced, including genes that previously were not known to be involved in the intermediate-early phase of fibroblast growth. Watching the process over time, the group assembled a data base of patterns that correspond with the cell's response to serum stimulation. By analogy, this system could be used to analyze gene-environment interactions, including looking for potentially meaningful polymorphisms.

## **Session VI: Ethical, Legal, Social Issues**

### **Introduction**

The Environmental Genome Project deals with environmental genetics; hence, its subject and activities involve a convergence of two socially charged areas—genetics and environmental health. Key issues will include informed consent and a wide range of

social, ethical and legal issues surrounding the design of epidemiological studies. If genetic traits identified in this project show that some individuals are more or less susceptible to certain agents, there could be profound implications in environmental regulation, and that might lead to more ethical quandaries. Powerful interests may line up in opposition to new regulations designed to protect the most susceptible members of society. Employers, for example, may find that instead of limiting exposures to harmful agents, it may be more convenient to discriminate against the genetically susceptible, since compliance with occupational regulations might be easier to achieve than compliance with environmental regulations.

This project will challenge the research community to be actively involved in social, ethical and legal discussion—and not leave the debate exclusively in the hands of lawyers, policy makers and professional ethicists. Leaving these matters to others will lead to overly restrictive legislation that will seriously impede research in this area. Or it will lead to flagrant abuses for which there may be reactive, regulative reform rather than reform based on sound science.

### **Ethical and Social Issues in Sampling**

The project's two phases present an array of issues that are not entirely scientific. Phase I will be the discovery of polymorphisms, including the various SNPs, with the greatest interest directed at those which affect phenotype. Phase I will involve a catalog, a reference set of DNA sequences, that one hopes will include many of the common variations in the human genome, with particularly good representation of those with functional consequences. Phase II is the design of studies that will attempt to correlate those changes in DNA sequence with particular phenotypes, that will collect the samples based on phenotypes of interest.

While there is overlap in the ethical, legal and social issues, they also are separable for these two phases, and the first part of this session deals with Phase I. The differences bear mentioning in a discussion about study design. The purpose of Phase I is not to understand the significance of alleles, but to find them. Phase II hopes to correlate significant alleles with phenotypes.

Bearing in mind that the aim of Phase I is discovery, there will be at least four issues to pay close attention to, and each of them come with pros and cons.

- Should individual identifiers remain on samples?
- Should any clinical information remain on samples?
- Should ethnic identifiers remain on samples?
- And should all genotypes for a particular DNA sample be collected in a common data base?

The first question, whether to leave identifiers on the samples, is probably the simplest to answer. There would be no reason to identify samples, unless participants want to know

the results. The question then becomes: Should the participants be denied that possibility? The most obvious reason for disallowing identification would be that if a lot of genotyping is done, sooner or later a deleterious marker will be identified in the DNA. If that information is tied to an individual and is a matter of public record, that person's identity can be discerned by anybody who wants to know who they are. That person, in fact, is at risk for a variety of dire consequences, including discrimination in health insurance and employment, and possible stigmatization. There's a practical issue, too. If data is going to go back to an individual, if results will be revealed that might eventually have an impact on that person's clinical management, then according to current regulations, all of the testing done on those DNA samples would have to be done in a certified laboratory. This would require every lab involved in genetic research on this set of DNA samples to go through an approval process, an onerous circumstance for the many research laboratories currently without such certification.

Should any clinical information remain with the samples? One could argue that one might get something for free here if common phenotypic information were recorded along with the DNA sample. Why not record whether individual participants are left-handed or bald? Or include blood pressure or I.Q.? Couldn't one deduce, on a genetic basis, some interesting phenotypes? Maybe, but would it be a good idea? That's not really the purpose of Phase I. For another thing, the likelihood of making incorrect inferences is quite substantial—there will be allelic differences between certain subsets, and one might conclude a cause and effect relationship on the basis of poor power or poor study design. Further, there's a risk of revealing the identity of the individuals. In some instances, sufficient information might be linked to a certain sample such that identification of the DNA donor becomes possible.

The remaining issues are thornier; the answers are less certain. Should ethnic identifiers remain on the samples? Should the sample even have ethnic representation? Clearly, if there are differences in populations in terms of allele frequencies, then sampling only people of northern European background is scientifically and politically a bad idea. So this reference sample should be diverse. Would that include sampling from other countries? Ultimately, if the goal of this U.S.-funded effort is to benefit human health of Americans, then perhaps a U.S. sample is entirely justifiable. But what minorities will be on the list? That's a significant question, as is whether to retain such information on each DNA sample. That might be useful information, might generate interesting information about the relatedness of various groups; about the degree of polymorphism in one ethnic group versus another. Further, people who are designing Phase II studies may want to know which particular ethnic group a particular allele is found in, and if they are looking at a SNP relevant to a particular disease, it would be useful to know what portion of the population that particular SNP was found in. If the portion corresponded to a particular ethnic group, they would probably want to conduct a follow-up study to determine the relationship to a particular phenotype. So there are reasonably strong scientific reasons to keep ethnic identifiers associated with samples.

But there are also strong arguments for disallowing ethnic identifiers on samples, chief among them the question of how to define “ethnicity.” Presumably, this would be self-identified—a method that is quite imperfect, as is the concept of ethnicity. Many people are predicting that, scientifically anyway, ethnicity will have less and less meaning. If ethnic identifiers are on this data set, it will be clear which groups were collected and which groups weren’t. If the identifiers are stripped off and the reference set represents the U.S. and has over-sampled minorities, it will be a little less obvious exactly which minorities were sampled and which were not. People may have questions. Why were Eskimos included? Or not included? People belonging to a particular ethnic group will be upset if they are included, and others belonging to the same group will be upset if they are excluded. Before the project goes much further, its planners must decide this issue of sampling ethnic groups and should seek input from the community. Perhaps most of the concern is in the potential for inadvertent stigmatization of an entire ethnic group. Imagine a study on a gene involved in alcohol metabolism. the gene may have nothing to do with the tendency toward alcoholism, but it is not difficult to picture an inflammatory headline like, “Researcher Discovers Gene that Explains Alcoholism in *a specified ethnic group.*” From the standpoint of possible social damage to entire groups whose members may not have consented to ethnic identifiers, it may be better to dispense with this information.

Finally, there is the question about whether genotypes from a DNA sample should be entered into a common data base. There are strong scientific reasons to consider doing this. For instance, if someone is studying gene A and if someone else is studying gene B and the genes happen to be 10 kb apart, and they deposit their data in the same data base, they may be able to jointly construct haplotypes that would be useful in understanding how that part of the chromosome has evolved. Without a common data base, the increased gain of knowledge would not be achieved. It also encourages a collaborative spirit. Everyone participating in this effort is linked to this common place to deposit data. This allows the possibility to cross-check for errors because the same people can check the same alleles on the same DNA sample and know that they should get the same answer, and if they don’t, they know there is a problem.

The obvious argument against doing this is as follows: The information associated with a given sample will uniquely identify it after several genotypes are known—this will essentially distinguish one person (and their DNA) from the other 6 billion people on the planet as the owner of this particular pattern of polymorphisms. But, to what will someone compare that person’s polymorphic signature? Will someone looking over data actually recognize, say, Jane Smith, and what difference could it make? It may be possible to learn about the abundance of a particular allele in this ethnic group versus that, and this common set of 100,000 or so in DNA samples makes a strong statement of likely ethnicity of the person from which that DNA sample came. It may lead to stigmatization of ethnic groups—not because, as in the earlier example. the ethnic identifiers have been retained, but because somebody figured out ethnicity later on. This is something to consider, but in this instance the scientific value of having such a data base is so high that having this option should be seriously considered.

There is another issue, and this is what constitutes informed consent for the Phase I study? What would a donor need to know when he or she donates a DNA or tissue sample for these studies? In the Human Genome Project, volunteers were solicited from locales with the capacity to make human reference sequence libraries. Before they donated DNA samples as either blood or sperm, they were put through an explicit informed consent process that included assurances of anonymity. A subset of those individuals were chosen for actual library construction so that not everybody who volunteered actually ended up in the sample. The identifiers were stripped off. The people who interacted with the volunteers were a different set than those working in the laboratories where the libraries were made. The DNA from which the human reference sequence will be determined is probably from a pool of fewer than 10 people. This might be a good model for Phase I because the issues are similar, the main concern being over uncovering something about subjects that might be used against them without their having been apprised of this possibility.

The standard guidelines on human subjects issues, including informed consent, typically dictate that as long as the sample is anonymous, consent can be waived. Still, it might be difficult arguing that anonymity could absolutely be guaranteed; for instance, it would be foolish to recommend that people go to the blood bank and grab some outdated blood off the shelf for DNA samples in Phase I of this variation project. Participants whose DNA will be used should have an opportunity to know when that will happen.

Keeping human-subject considerations in mind, is there a way to jump-start Phase I? Possibly, through the CDC's health and nutrition survey. A part of the project involved the collection of clinical information and DNA samples and other biological materials from 20,000 individuals. The study was set up in such a way that it purposely over-sampled minorities, particularly African-Americans and Hispanics. There are 8,000 cell lines already available from individuals who participated in this study, and an institutional review board has suggested it would be acceptable for those samples to be used for allele frequency determinations so long as the identifiers are stripped off. It is not clear whether reviewers would be comfortable, however, with having a common data base into which all allele information might be submitted on a particular sample. It would be advantageous if such a sample could be also used in the Environmental Genome Project, at least as part of the Phase I data set. The chance of using this data for trios is slim. Most of these samples were not collected as families but as individuals.

Finally, the design of this Phase I DNA resource has implications for all of biomedical research. This set of DNA samples will be used by the scientific community for a very large number of applications. To get it right, there must be broad input from scientists and from non-scientists. There has not been enough input from the latter group, particularly on the topic of inclusion or non-inclusion of various ethnic groups into the sample set and whether or not it will be acceptable to leave ethnic identifiers on the samples.

## **Informed Consent, Potential Impact on Susceptible Groups**

Over the next few years, as the Environmental Genome Project attempts to assemble sample banks that combine DNA and environmental information, questions concerning the associated ethical issues, such as what constitutes appropriate informed consent, will confront everyone who is involved.

One of the major issues will be waivers. Waivers can be open to many interpretations. An investigator could ask to be waived from the requirement of explicit consent for genetic research if that research involves no more than minimal risk. The obvious problem here is that one person's minimal risk might be someone else's moderate risk, and it's very difficult to define that concept. Can the research be carried out without the waiver? This actually has a substantial impact on epidemiological studies because it will be difficult to get explicit, detailed consent with the large sample sizes this project will entail. So the question of what is practical depends on how much funding is available for genetic counseling and pre- and post-test counseling. Wherever appropriate, the subjects will be provided with additional pertinent information after participation.

Now the big question becomes informed consent to what? If an investigator is interested in a particular class of genes, for instance, perhaps he or she would ask for consent to use only those genes—mismatch repair genes, or something more specific. Some ethicists maintain consent cannot be informed unless the gene of interest has been specified, unless something is said about the pros and cons of being analyzed for that particular gene. That, obviously, would be a very onerous task, especially if one were interested in multiple genes. Some have even proposed that specific mutations in individual genes carry different risks and different implications, so perhaps counsel should be offered at the mutation-specific level.

Will a generic consent cover all genes? Will consent be assumed for 50 or 60 genes and/or polymorphisms of genes on a single chip? In a prospective study, these will be big questions. When people enroll in a study, nobody knows who among them will get cancer or cardiovascular disease or Alzheimer's or diabetes. Some propose that it is not appropriate to ask for blanket consent for every disease, that subjects should check off diseases on a list and specify whether they would consent to be in future studies of cancer or cardiovascular disease but might prefer not to be in a study of alcoholism or a psychiatric condition and other more socially loaded maladies.

A subject may suffer considerable anxiety upon learning of a predisposition to a particular disease or syndrome, particularly if there is no known intervention. The most obvious risk here is health insurance discrimination, but employment discrimination is a major concern as well. There is the possibility that individuals uniquely susceptible to carcinogens or toxins in a particular industrial environment will be screened out of employment on the pretext that companies were looking out for the welfare of the workers. It would also raise the possibility that employers and polluters could be more lax by claiming to pose no harm when exposing resistant populations. One survey

suggested that by the year 2000, 20 percent of employers would be conducting some sort of genetic screening, so this will be a lively issue in the years to come. This screening might also be applied to susceptible subgroups. Susceptibility may be defined on the basis of ethnicity or some other grouping, and that is clearly a social consequence that should raise concerns.

While people are nervous about the misapplication and misuse of genetic testing, there are potential benefits to a person's knowing his or her genetic susceptibility results. This raises the question of whether investigators have a duty to inform people in their studies who might benefit from this information. In some cases, knowledge of predisposition might reduce anxiety. Such knowledge can lead to informed decisions about specific lifestyle change—people can remove things from their environment like cigarette smoke and harmful foods, and they can remove themselves from harmful environments, and by taking action they might reduce their risk of disease. One example from this meeting showed that people susceptible to esophageal cancer had much lower mortality from the disease when they were frequently screened. In the Environmental Genome Project, what should study participants know? What are the criteria for giving feedback? Does one do feedback on all data? On all those green and yellow lights on a DNA chip? Obviously not. What about clinically relevant data? And who will decide what is clinically relevant? Preliminary results? The first time an association is made?

A series of logistical constraints must be incorporated into this discussion. New consents may have to be obtained from people already enrolled in studies if they didn't give the standard consent for future studies. As in any epidemiological study, many people will be lost to follow-up in case-control studies. It would be enormously difficult to find cases from a case-control study conducted 10 years earlier. People may be dead, and there is controversy about that—what steps would one follow to genotype a sample from a deceased person? Permission from next of kin? If so, how would that work?

Another big problem is that large studies tend to be far-flung. Patients in the study may be spread across an entire state or an entire region or maybe even the entire country, and they may not check in routinely, as they would in a typical clinical study.

Finally, there is federal and state legislation pending that is designed to prevent health insurance discrimination. One way to do that would be to prohibit genotyping without specific informed consent. That, obviously, would have major implications for research. Different bills vary in how they target insurance discrimination; one draft of a bill would ask any investigator who holds a DNA sample (or any sample that could be used to extract DNA) to obtain a subject's consent—not for further research but to simply hold that sample. If consent were not obtained, the sample would be destroyed. That, obviously, would have a major impact on samples now in banks.

One-size-fits-all regulations or legislation cannot accommodate all of the complexities and differences in genes and populations and in study designs involved. It would be problematic if federal legislation passed that was ultra-rigorous on some of these issues,

because it probably could not accommodate the complexity of the different studies and populations involved. A model for solving these problems is the same one to regulate all other aspects of human health research: the institutional review board. IRBs around the country deal with enormously tricky issues and they can deal with this one, though human subjects committees need much greater education, better guidelines and much more input on accepted community standards with respect to genetic epidemiological research. The challenge for the Environmental Genome Project is to design informed consent procedures that are both feasible and ethically unimpeachable.

## PROGRAM PARTICIPANTS

**J. Carl Barrett**

Division of Intramural Research  
National Institute of Environmental Health Sciences, NIH  
P.O. Box 12233  
Research Triangle Park, North Carolina 27709-2233

**Douglas A. Bell**

Division of Intramural Research  
National Institute of Environmental Health Sciences, NIH  
P.O. Box 12233  
Research Triangle Park, North Carolina 27709-2233

**Patrick O. Brown**

Department of Biochemistry, HHMI  
Stanford University Medical Center  
255B Beckman Center  
Stanford, California 94305-5428

**Kenneth H. Buetow**

Division of Population Sciences  
Fox Chase Cancer Center  
7701 Burholme Avenue  
Philadelphia, Pennsylvania 19111

**David C. Christiani**

Harvard School of Medicine & Public Health  
665 Huntington Avenue  
Boston, Massachusetts 02115

**Francis S. Collins**

National Human Genome Research Institute, NIH  
31 Center Drive, MSC 2152, Room 4B-09  
Bethesda, Maryland 20892-2152

**David R. Cox**

Genetics Department  
Stanford University Medical Center  
Stanford University  
Stanford, California 94305-5120

**Maureen T. Cronin**  
Affymetrix, Inc.  
3380 Central Expressway  
Santa Clara, California 95051

**Georgia M. Dunston**  
Department of Microbiology  
Howard University  
College of Medicine  
520 W. Street, N.W.  
Washington, DC 20059

**Glen A. Evans**  
McDermott Center for Human Growth & Development  
UT Southwestern Medical Center  
Simmons Biomedical Research Building  
5323 Harry Hines Boulevard  
Dallas, Texas 75235-8591

**Joseph F. Fraumeni**  
Division of Cancer Epidemiology and Genetics  
National Cancer Institute, NIH  
Bethesda, Maryland 20892

**Frederick P. Guengerich**  
Department of Biochemistry  
Center in Environmental Toxicology  
Vanderbilt University School of Medicine  
Nashville, Tennessee 37232-0146

**Dean H. Hamer**  
Laboratory of Biochemistry  
National Cancer Institute, NIH  
Bethesda, Maryland 20892-0001

**Susan E. Hankinson**  
Channing Laboratory  
181 Longwood Avenue  
Boston, Massachusetts 02115

**Leland H. Hartwell**  
Fred Hutchinson Cancer Research Center  
1124 Columbia Street  
Seattle, Washington 98104