# The Environmental Genome Project: Functional Analysis of Polymorphisms

*F. Peter Guengerich*

Department of Biochemistry and Center in Molecular Toxicology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232-0146 USA

The ultimate goal of the Environmental Genome Project is the improvement of human health on the basis of information about the variations in certain genes. The first phase of the project will involve the selection of the human genes and characterization of the alleles occurring in the American population. However, intelligent use of this information will require analysis of the relevance of the allelic differences. Epidemiology alone will not solve the problem, and mechanistic studies will be required. Factors to be considered in the design of functional analyses are considered in this commentary. *Key words*: analysis of function in genes, cytochrome P450, Environmental Genome Project, enzymes involved in metabolism, expression of proteins, functional polymorphisms, human genome, limitations of epidemiology, polymorphisms. *Environ Health Perspect* 106:365–368 (1998). [Online 1 June 1998]
*http://ehpnet1.niehs.nih.gov/docs/1998/106p365-368guengerich/abstract.html*

The Environmental Genome Project is an effort by the NIEHS to characterize the variations in a number of important human genes and to relate these differences to susceptibility of humans to chemical and physical agents in the environment. The long-term goal of this work is to improve human health, primarily through improved prevention, but also through treatment when appropriate.

Why is there strong interest in this project? At least four major reasons can be identified. 1) The Human Genome Project is not finished, but there is broad support in the biomedical science community for the program and a general consensus that the work will be completed on schedule, even in the absence of major new developments in technology. Also, related government efforts are under way, such as the Centers for Disease Control National Health and Nutrition Examination Survey (NHANES) program and a National Cancer Institute effort to identify genes that show altered expression levels in tumors. All of this information will provide a knowledge base to use in disease prevention and treatment. 2) There is a long history of interest in inherited diseases, which resulted in the establishment of medical genetics and genetic epideminology as disciplines (*1*). Many of these debilitating diseases were found to be disorders of primary metabolism. The reasoning is that if genetics contributes to rare diseases, then it should also contribute to more common ones. 3) Many studies with experimental animals show genetic influences on susceptibility to chemical and physical agents. For instance, mice deficient in the *Ah* locus show differences in cancer incidence and acute toxicity from a variety of chemicals (*2*). Another relevant example in humans is the differential susceptibility to ultraviolet (UV) light due to functional differences in DNA repair genes (*3,4*). 4) Studies on human drug

metabolism over the past 20 years have provided convincing evidence of the wide variability in function of the enzymes of xenobiotic metabolism [e.g., cytochromes P450, N-acetyltransferases, and glutathione (GSH) transferases] and the *in vivo* significance of these variations in the disposition of these drugs in humans (*5–7*). In some cases, fatalities can be attributed to these differences (*8*). Further, many of these same enzymes are known to be involved in the bioactivation and detoxication of potentially health-damaging chemicals found in the environment (i.e., other than drugs).

## Background

The goal of the Environmental Genome Project is to characterize allelic variants qualitatively and quantitatively. Variations from a predominant allele are often referred to as genetic polymorphisms, a term used to describe variants occurring at an incidence of >1%. Polymorphisms are common among humans. Some have no functional significance, and identification of which do and do not have functional significance will be necessary in order to interpret the information obtained in the Environmental Genome Project.

The significance of a polymorphism depends on the phenotype. Phenotype changes are observed if 1) the altered codon produces an amino acid substitution that changes the catalytic or stability properties of the coded protein; 2) a stop codon is formed to cause premature termination of protein synthesis; 3) the rate of transcription to form RNA is decreased or increased; 4) RNA maturation is prevented (splice site variants); or 5) mRNA stability is modified. Only case 1 produces a protein with different properties, but any of these cases can raise or lower the amount of a particular protein produced.

Two other points should be made. First, there are emerging technologies that can be used to rapidly assay mRNA levels. At the recent NIEHS Symposium on the Environmental Genome (*9*), P.O. Brown presented information on the use of chip technology for this purpose (*10*). Nongenetic effects on levels of mRNA and protein expression cannot be distinguished from genetic effects in the absence of further information. However, if a genetic polymorphism is identified, it may be possible to use mRNA expression to establish the functionality of a polymorphism.

## Use of Epidemiology: Strengths and Weaknesses

Considerable effort has been expended on the use of epidemiology in the analysis of the effects of physical and chemical agents in the environment on human health. Recently, the limitations of epidemiology have been discussed (*11,12*). Inherent problems in most epidemiological analysis include racial patterns, small sample sizes, and confounding variables. For these reasons, it is unlikely that epidemiology alone will be sufficient to establish the significance of polymorphisms. An analogy, although not perfect, is in drug development, where clinical trials would not be done in the absence of at least some basic mechanistic information about the action of a new drug.

The environmental health epidemiology literature contains many controversies. The effects of low-level electromagnetic radiation on cancer and pesticides on breast cancer remain controversial at best (*11,13–16*). The field of molecular epidemiology, in which biomarkers and intermediate end points are incorporated, is also complicated. Many associations of genetic polymorphisms of enzymes involved in xenobiotic metabolism with disease have been considered (*17*), and meta-analyses are presented in a recent article by d'Errico et al. (*18*).

Odds ratios vary considerably among different studies with some of the genes. Further, in many cases there is no established mechanistic basis for associations. Although the Ile/Val 462 polymorphism of P450 1A1 has been studied extensively, two studies have shown little if any catalytic differences between the gene products (19,20). A reported association of an epoxide hydrolase polymorphism with aflatoxin-related hepatocellular carcinoma (21) has also been questioned on a mechanistic basis (22).

Although these are serious concerns about the discipline of classical epidemiology, ultimately mechanism-based molecular epidemiology will be needed to establish the importance of paradigms developed in the laboratory. Returning to the analogy of drug development, many candidate compounds appear promising, but do not deliver efficacy in clinical trials.

## Problems in Functional Analysis

As mentioned above, there are serious caveats in reliance on only epidemiology as a means to establishing significance of genetic variations. One reasonable approach would be to use experimental studies to establish mechanisms, which can be used as a guide to doing more enlightened epidemiology. However, there are serious issues to consider in designing meaningful mechanistic studies on the functional significance of polymorphisms.

Many of the environmental exposures are complex mixtures, e.g., tobacco smoke, smog, and mixtures of solvents in waste dumps. The etiology is often not precisely known. Many of these mixtures are not readily adaptable to laboratory experiments (23–25). Another issue is the choice of an appropriate dose of a single compound or a mixture in a cellular system. What dose is really relevant to the human disease under consideration? Will this dose be enough to observe a difference in the laboratory experiment? Is the end point under observation reflective of the disease? For instance, DNA adducts and mutations are often used as intermediate end points in cancer studies but may not always be predictive in the absence of other factors such as cell proliferation. We also need to know whether a certain assay will be predictive for the target organ under consideration.

Another potential problem is an unexpected gain of an unpredicted function. This appears to be the case in superoxide dismutase variations related to familial amyotropic lateral schlerosis, in which the variant form appears to enhance $H_2O_2$-dependent oxidations that may be responsible for the neurological defect (26,27).

Similarily, a certain P450 variation might show decreased catalytic activity toward a model substrate but increased activation of one more relevant to toxicity, and the effect might not be predicted.

Finally, another complication is the interactions of gene products, an issue related to haplotypes. This situation can be readily appreciated with some of the proteins involved in gene regulation, in which various heterodimers have varying activities and can be derived from members of two different gene families, e.g., PPAR/RXR and fos/jun (28,29). Such situations can also be seen with enzymes involved in xenobiotic metabolism, in which one enzyme makes a product that is used as a substrate by another. Such paradigms apply to pathways of both activation and detoxication.

## Approaches to the NIEHS Environmental Genome Project

Although the outline of the Environmental Genome Project was still being developed at the time of the October 1997 workshop, the general strategy being developed is as follows. The first stage is defining the precise goals, genes under investigation, numbers and source(s) of samples, and the technologies to be employed in the sequencing analysis. The first real experimental stage is the actual sequence analysis. Currently a rough estimate is that this will involve 200 genes and 1,000 human samples, and if we assume an average of 10 kb/gene then the prospect is for approximately 2 billion bases of sequence analysis (9).

The next stages are less clear (9). One aspect involves extending the basis to more individual samples. This can readily be accomplished using chip technology, in the context of all of the identified polymorphisms. The other aspect involves functional analysis. Knowledge of the polymorphic variants will be of some immediate use in epidemiological settings, although the concerns about understanding significance of polymorphisms through only epidemiological studies have already been expressed. Functional analysis must involve mechanistic studies on the roles of the polymorphisms. However, exactly how the studies will be designed has not been considered in detail as of yet.

## Strategies for Functional Analysis

In some cases assays can be done directly with humans. For instance, drugs are oxidized by some of the P450 enzymes under consideration here (5). Function can often be determined in in vitro assays in tissues obtained from organ donors or surgical

samples, and comparisons with genotype can be made. Such work can be extended to in vivo drug studies with humans (30). Thus, information about polymorphisms that will be relevant to drug metabolism itself can be rapidly obtained. The question is how to relate this to diseases with complex etiologies (e.g., chemical mixtures, unknown causes).

Another approach, which will probably be more general, involves heterologous expression and examination of function. There are basically two different end points to use. One is to use a specific assay (e.g., for an enzyme such as glutathione peroxidase); the other is to examine the pathology of the cells following expression of the polymorphic variants of each gene product. Questions involve appropriateness of these model systems to humans.

A number of model systems exist for heterologous expression of genes (31). They include bacteria, insect cells, yeast, mammalian cells, human cells, and (transgenic) mice. Human cell lines are usually of tumor origin and may either involve transient expression or be stable cell lines. A number of issues need to be considered when choosing a system. They include ease of expression, yields of expressed product, the endogenous background (i.e., similar proteins to the expressed gene product), the availability of discernible parameters as end points to measure, and relevance of the system to the human situation. The similarity to humans must be considered with regard to transcriptional regulation, protein stability, and interactions with other components.

## Approaches to Functional Analysis—A Universal Approach

There are two approaches to doing functional mechanistic analysis of polymorphisms. One strategy is to choose a common cellular system and express all of the genes under investigation. This has the advantage of a common base for comparison and could also make use of the cheapest host system. A possible proposal is to use yeast to express all of the genes and their variants. Yeast are eukaryotes and have an advantage in that many, but not all, of the gene products involved in yeast gene regulation have been found to have human counterparts.

Following expression of genes, the enzyme activity could be measured, if an assay exists. The phenotypic changes could be observed and quantified, if possible. These results could then be extended to (human) epidemiology studies.

There are a number of deficiencies in this approach. The main one is that the

assays of function will be quite different for the various genes, and some may not be appropriate for yeast. For instance, the enzymes involved in xenobiotic metabolism tend to act in concerted pathways, and yeast do not normally contain much of these systems.

## Multiple Approaches to Functional Analysis: Examples

An approach that will probably be more realistic is one that involves multiple strategies depending upon the gene under consideration. A few examples are considered.

In many respects, the enzymes involved in the metabolism of xenobiotic chemicals are easiest to deal with. After establishing the polymorphisms and frequency distribution, it may be expedient to first look for a correlation of the polymorphism and *in vitro* catalytic activity in human tissue samples. Various gene products can also be expressed readily in simple heterologous vector systems, and assays of catalytic activity can be done. These enzyme systems have the advantage that, in many cases, probe drug substrates are known that can be used in humans (*5*), and studies on *in vivo* phenotyping/genotyping may be possible, as already have been done for enzymes such as P450s 2D6 and 2C19 (*32*). If differences can be identified, work can be extended to epidemiological settings. A caveat in all of this is that the test (and probe) substrates may not be relevant to human health issues, so aspects such as the chemical etiology of disease must also be considered.

Another case involves enzymes related to oxidative stress. As with the case above, it is possible to look for correlation of the polymorphism and *in vitro* catalytic activity in human tissue samples, if relevant samples and an appropriate assay are available. The gene products can be expressed in simple systems, perhaps even microbial ones, and the abilities of the expressed gene products can be compared with regard to their abilities to protect from an induced oxidative stress. If microorganisms are used, some of the experiments might be repeated in mammalian cells. In the extension of the work to molecular epidemiology, one approach might be to look for biomarkers of oxidative stress [e.g., $F_2$ isoprostanes (*33*)] in humans having various genotypes.

Enzymes that repair chemical and physical damage to DNA are certainly of interest in a venture such as the Environmental Genome Project. One approach might be to use blood cells from individuals of defined genotypes and do assays on the repair of transfected plasmids with UV- or carcinogen-damaged DNA (*34*). The work

could then be extended to cell systems in which the individual gene products are overexpressed and then treated with chemicals or physical agents known to produce the particular kind of DNA damage under consideration. The results could then be extended to epidemiology.

Some of the systems under consideration are complex, and more creativity may be required. For instance, how will we deal with polymerases, tumor suppressors, cell cycle checkpoint-associated proteins, and signal transduction pathways? In some cases, coupling may be done in yeast experimental systems. A general problem with both yeast and mammalian cells is that wild-type homologs already exist in the cells and may make interpretation of expression studies more difficult.

## Conclusions

A few more points should be made. In the above considerations of the possible expression systems to use, one might ask why not use the most complex and relevant. However, if we work with 200 genes and each has 5–10 polymorphisms (a rough but not unreasonable estimate), we face the prospect of 1,000–2,000 expressions and doing the necessary assays. Therefore, some thought must be given to simplicity of systems and rapid assays.

Another problem already mentioned is that many of the effects may be small in themselves, and synergism of multiple phenotypes may be needed for a particular disease. For instance, one can consider the epidemiology work on colon cancer, where a mixture of a high level of P450 1A2 and the rapid phenotype for *N*-acetyltransferase 2 showed a synergism (along with consumption of well-done cooked beef) (*35*). Further, we do not really know all of the functions of the genes we are exploring, due to various aspects of coupling. For instance, some proteins appear to be involved in both DNA repair and transcription (*36–38*).

Finally, we have issues of inducibility and various nongenetic differences. In many cases the regulatory elements involved have not been identified, and these should be the targets for polymorphism studies. Also, we should remember that unknown dietary factors play an important role in these issues of human health, and the genotyping studies may or may not impact on these.

Despite all the caveats presented about how to proceed with functional analysis, there is still considerable optimism that the Environmental Genome Project will be successful. Accumulation of genotype data is clearly the first step. Planning is critical for the functional analysis stage.

### REFERENCES AND NOTES

1. Scriver CR, Beaudet AL, Sly WS, Valle D, eds. The Metabolic and Molecular Bases of Inherited Disease. 7th ed. New York:McGraw-Hill, 1995.
2. Nebert DW. The *Ah* locus: genetic differences in toxicity, cancer, mutation, and birth defects. Crit Rev Toxicol 20:153–174 (1989).
3. Cleaver JE, Kraemer KH. Xeroderma pigmentosum. In: The Metabolic Basis of Inherited Disease, Vol 2 (Scriver CR, Beaudet AL, Sly WS, Valle D, eds). 6th ed. New York:McGraw-Hill, 1989;2949–2971.
4. Friedberg EC. Xeroderma pigmentosum, Cockayne's syndrome, helicases, and DNA repair: what's the relationship? Cell 71:887–889 (1992).
5. Guengerich FP. Human cytochrome P450 enzymes. In: Cytochrome P450 (Ortiz de Montellano PR, ed). New York:Plenum Press, 1995;473–535.
6. Guengerich FP, ed. Biotransformation. Comprehensive Toxicology, Vol 3. Oxford:Elsevier-Pergamon, 1997.
7. Mahgoub A, Idle JR, Dring LG, Lancaster R, Smith RL. Polymorphic hydroxylation of debrisoquine in man. Lancet 2:584–586 (1977).
8. Guengerich FP. Role of cytochrome P450 in drug interactions. Adv Pharmacol 43:7–35 (1997).
9. Kaiser J. Environment Institute lays plans for gene hunt. Science 278:569–570 (1997).
10. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686 (1997).
11. Taubes G. Epidemiology faces its limits. Science 269:164–169 (1995).
12. Shy CM. The failure of academic epidemiology: witness for the prosecution. Am J Epidemiol 145:479–484 (1997).
13. Wolff MS, Weston A. Breast cancer risk and environmental exposures. Environ Health Perspect 105(suppl 4):891–896 (1997).
14. Ahlborg UG, Lipworth L, Titus-Ernstoff L, Hsieh CC, Hanberg A, Baron J, Trichopoulos D, Adami HO. Organochlorine compounds in relation to breast cancer, endometrial cancer, and endometriosis: an assessment of the biological and epidemiological evidence. Crit Rev Toxicol 25:463–531 (1995).
15. Krieger N, Wolff MS, Hiatt RA, Rivera M, Vogelman J, Orentreich N. Breast cancer and serum organochlorines: a prospective study among white, black, and Asian women. J Natl Cancer Inst 86:589–599 (1994).
16. Dewailly É, Dodin S, Verreault R, Ayotte P, Sauvé L, Morin J, Brisson J. High organochlorine body burden in women with estrogen receptor-positive breast cancer. J Natl Cancer Inst 86:232–234 (1994).
17. Perera FP. Environment and cancer: who are susceptible? Science 278:1068–1073 (1997).
18. d'Errico A, Taioli E, Chen X, Vineis P. Genetic metabolic polymorphisms and the risk of cancer: a review of the literature. Biomarkers 1:149–173 (1996).
19. Zhang Z-Y, Fasco MJ, Huang L, Guengerich FP, Kaminsky LS. Characterization of purified recombinant human CYP 1A1-Ile$^{462}$ and Val$^{462}$: assessment of a role for the rare allele in carcinogenesis. Cancer Res 56:3926–3933 (1996).
20. Person I, Johansson I, Ingelman-Sundberg M. *In vitro* kinetics of two human *CYP1A1* variant enzymes suggested to be associated with interindividual differences in cancer susceptibility. Biochem Biophys Res Commun 231:227–230 (1997).
21. McGlynn KA, Rosvold EA, Lustbader ED, Hu Y, Clapper ML, Zhou T, Wild CP, Xia X-L, Baffoe-Bonnie A, Ofori-Adjei D, et al. Susceptibility to hepatocellular carcinoma is associated with genetic variation in the enzymatic detoxification of aflatoxin $B_1$. Proc Natl Acad Sci USA 92:2384–2387 (1995).
22. Johnson WW, Ueng Y-F, Yamazaki H, Shimada T, Guengerich FP. Role of microsomal epoxide hydrolase in the hydrolysis of aflatoxin $B_1$ 8,9-epoxide. Chem Res Toxicol 10:672–676 (1997).
23. Mauderly JL. Toxicological approaches to complex mixtures. Environ Health Perspect 101(suppl 4):155–165 (1993).
24. Shimada T, Guengerich FP. Activation of amino-α-carboline, 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine, and a copper phthalocyanine cellulose

extract of cigarette smoke condensate by cytochrome P-450 enzymes in rat and human liver microsomes. Cancer Res 51:5284–5291 (1991).

25. McCann J, Dietrich F, Rafferty C, Martin AO. A critical review of the genotoxic potential of electric and magnetic fields. Mutat Res 297:61–95 (1993).

26. Wiedau-Pazos M, Goto JJ, Rabizadeh S, Gralla EB, Roe JA, Lee MK, Valentine JS, Bredesen DE. Altered reactivity of superoxide dismutase in familial amyotrophic lateral sclerosis. Science 271:515–518 (1996).

27. Tu PH, Raju P, Robinson KA, Gurney ME, Trojanowski JQ, Lee VMY. Transgenic mice carrying a human mutant superoxide dismutase transgene develop neuronal cytoskeletal pathology resembling human amyotrophic lateral sclerosis lesions. Proc Natl Acad Sci USA 93:3155–3160 (1996).

28. Johnson EF, Palmer CNA, Griffin KJ, Hsu MH. Role of the peroxisome proliferator-activated receptor in cytochrome P450 4A gene regulation. FASEB J 10:1241–1248 (1996).

29. Angel P, Hattori K, Smeal T, Karin M. The jun proto-oncogene is positively autoregulated by its product Jun/AP-1. Cell 55:875–885 (1988).

30. Rettie AE, Wienkers LC, Gonzalez FJ, Trager WF, Korzekwa KR. Impaired (S)-warfarin metabolism catalysed by the R144C allelic variant of CYP2C9. Pharmacogenetics 4:39–42 (1994).

31. Goeddel DW, ed. Methods in Enzymology, Vol 185, Gene Expression Technology. San Diego, CA: Academic Press, 1990.

32. Meyer UA, Zanger UM. Molecular mechanisms of genetic polymorphisms of drug metabolism. Annu Rev Pharmacol Toxicol 37:269–296 (1997).

33. Morrow JD, Roberts LJ II. The isoprostanes: current knowledge and directions for future research. Biochem Pharmacol 51:1–9 (1996).

34. Athas WF, Hedayati M, Matanoski GM, Farmer ER, Grossman L. Development and field-test validation of an assay for DNA repair in circulating human lymphocytes. Cancer Res 51:5786–5793 (1991).

35. Lang NP, Butler MA, Massengill J, Lawson M, Stotts RC, Hauer-Jensen M, Kadlubar FF. Rapid metabolic phenotypes for acetyltransferase and cytochrome P4501A2 and putative exposure to food-borne heterocyclic amines increase the risk for colorectal cancer or polyps. Cancer Epidemiol Biomarkers Prev 3:675–682 (1994).

36. Hanawalt PC. Role of transcription-coupled DNA repair in susceptibility to environmental carcinogenesis. Environ Health Perspect 104(suppl 3):547–551 (1996).

37. Sancar A. DNA excision repair. Annu Rev Biochem 65:43–81 (1996).

38. Friedberg EC. Relationships between DNA repair and transcription. Annu Rev Biochem 65:15–42 (1996).