Metabolic Polymorphisms and Susceptibility to Cancer
W. Ryder
IARC Scientific Publications No. 148
International Agency for Research on Cancer, Lyon, 1999

Chapter 4. Selection of candidate genes for population studies

Neil Caporaso

A broad view of genetic susceptibility in humans suggests that high-penetrance hereditary genes cause a number of relatively uncommon tumours in the familial setting, while common cancers are influenced by multiple susceptibility loci. Early investigations of the latter category focused on the role of genes in the metabolism of carcinogens (activation, detoxification) while current and planned studies extend to genes with diverse mechanisms involving DNA repair, cell cycle control, nutrient metabolism and other processes. The present report considers some methodological issues pertinent to the study of the common genes, focusing in particular on the selection of appropriate candidates for study.

A central issue in population studies that consider putative genetic risk factors is the selection of appropriate candidate genes for study. The 'susceptibility genes' of interest in this connection overlap but differ in many respects from the 'human disease genes' or high penetrance genes that have been identified for a variety of rare cancers and a subset of common cancers. Typically, the high penetrance disease genes are uncommon (i.e. have a low gene frequency, typically less or much less than 1%), when present confer a high relative and absolute risk, and have been identified and studied in the family setting (Caporaso & Goldstein, 1995) (Table 1).

Two general approaches have been used to identify candidate genes potentially involved in familial cancers: the functional approach and the positional approach. In the *functional approach*, used in many early studies, a known biochemical basis for pathogenesis is exploited to generate a gene product or to perform functional complementation cloning. For example, the addition of certain fragments of human chromosomes to cell lines from individuals who have various DNA repair defects can produce a repair competent phenotype (Auerbach *et al.*, 1998). This method led to the identification of Fanconi's anaemia. Pharmacogenetic approaches are related in that heritable differences in enzymes thought to have a function relevant to the condition of interest are selected as candidates. A gene product that is known to control a metabolic transformation crucial to carcinogenesis, such as activation of benzo(a)pyrene by CYP1A1 or the Ah receptor, would draw attention to the gene for study in relation to susceptibility to cancer where that agent is a known risk factor.

More commonly the 'forward genetics' approach is not applicable since a responsible metabolic pathway is not known. Particularly for

Table 1. Genes: single and susceptibility

Factor	Single	Susceptibility
Gene frequency	Rare	Common (>1%)
Study setting	Family	Population
Study type	Linkage	Association
Penetrance	High	Low
Absolute/relative risk	High	Low
Population-attributable risk	Low	High
Role of environment	Modest	Critical

the rare cancers that exhibit Mendelian patterns of inheritance, 'reverse genetics' *positional approaches* have achieved notable successes. The term 'positional' refers to their dependence on the chromosomal location of the gene.

Positional strategies fall loosely into three categories. Linkage studies are an increasingly common method of identifying disease loci. This approach relies on locating genes through recognizing their cosegregation with marker alleles. This method requires a sufficient number of well-characterized families with the disease of interest, biological specimens (i.e. germline DNA from affected and unaffected individuals in the kindreds, a postulated genetic mechanism (i.e. autosomal dominant/recessive), a set of polymorphic markers, and a statistical approach to detect evidence for linkage (i.e. elevated LOD scores (LOD = log of odds; a value over 3 is generally accepted as evidence). A second method used to identify tumour-suppressor genes involves the detection of loss of heterozygosity. In this approach, paired normal and tumour tissues are examined for polymorphic markers. The loss of signal in target (usually, but not always tumour) tissue compared to normal tissue, if consistently observed in a particular chromosome region, is taken as evidence that a 'second hit' has occurred at a critical region representing an important gene (Wistuba *et al.*, 1997). Finally, the identification of chromosome abnormalities in tumour tissue may implicate a particular area (Nowell & Hungerford, 1960).

Once a subchromosomal region is identified by one of these approaches, further work is required to identify particular candidate genes within the area. Both the increasing repertoire of molecular approaches and the number of human genes that are mapped have rendered this stage of the process more tractable. Once identified, evidence to implicate a gene in pathogenesis can involve a variety of approaches: mutation screening (identifying mutations of the genes in individuals with the condition in question), demonstration of restoration of a normal phenotype when transfection of the cloned normal allele is introduced into a cell line with the mutant phenotype, or production of a mouse model of the disease using gene targeting to introduce the defective mutation.

The present paper considers aspects of the study of the more common, low penetrance genes postulated to play a role in many or even most common diseases. These genes are characterized by high allele frequency (1% to 90% or more), low relative and absolute risk, but potentially high population-attributable risk (Caporaso & Goldstein, 1995). Can the methods used to identify the family cancer genes be applied to the study of the susceptibility genes? The answer is mixed. Linkage analysis approaches face power constraints when the genotypic relative risk (GRR) is small. Risch & Merikangas (1996) observed that even under ideal conditions, i.e. a closely linked marker locus that is highly informative, the number of families required to detect a GRR of 2 would be around 2500, which would be prohibitive. Thus approaches based on association are likely to be required to examine candidate gene hypotheses involving low to modest relative risks. Hybrid approaches such as 'association mapping', exploiting linkage disequilibrium to identify candidate areas, require detailed genetic maps (thousands of evenly spaced genetic markers) and automated techniques to characterize polymorphic markers in pooled specimens. This approach, which should become feasible in the next few years, is likely to have improved power over linkage approaches (Lander & Kruglyak, 1995; Barcellos *et al.*, 1997).

The steadily increasing list of known genes, the improved technical means of extracting genetic information, and the diminishing quantities of biological material required to read the genetic record permit expanding opportunities to perform DNA studies. Automated or chip-based genotype recognition systems can be expected to further facilitate the acquisition of genetic information.

It is worth reflecting on the historical context of studies on susceptibility genes to better understand both the general direction of the field and the implications for newer studies. Only a decade ago, early studies of this type required laborious phenotype determinations in order to establish the innate activity of one gene. Examples include the work of Lower *et al.* (1979) on the acetylation phenotype and bladder cancer in workers occupationally exposed to aryl amines, the study of

Kellerman *et al.* (1973) on aryl hydrocarbon hydroxylase inducibility in lung cancer, and the work of Ayesh *et al.* (1984) on debrisoquine metabolism in lung cancer. In each of these studies, the effort required for phenotype determination imposed methodological limitations. For example, it was difficult and expensive to carry out phenotyping on substantial numbers of subjects, and bias was possible through exclusion of the sickest subjects who did not meet clinical criteria. A distinct advantage of the phenotyping approach is that the functional nature of the gene is established by the phenotyping operation. In theory this approach avoids the problem of absent underlying functional genetic variation or of mutations being too numerous or diverse to identify by genotyping. Some of the contrasts between the phenotyping era, the early genotyping phase, and the anticipated future course of studies are presented in Table 2.

In view of the new opportunities offered by current technologies, some discussions of criteria for gene selection are worth considering. This is particularly relevant, as current technologies offer enhanced ability to test multiple genes. There are pitfalls, however, notably a predictable increase in false positives due to multiple comparisons. This implies that function and biological plausibility are increasingly critical in guiding the interpretation of findings.

Table 2. Historical evolution of the study of common genes and cancer

	Phenotyping	Early genotyping	Future genotyping
Period	1974-1987	1987-present	Future
Number of genes	One	One to a few	Many per study
Type of genes	Metabolic polymorphisms (phenotype)	Metabolic polymorphisms (genotype)	Wide variety of gene categories (genotype with enhanced functional information)
Study design	Case and control series	Case-control	Cohort
Analysis	Main effect of gene	Main effect of gene	Gene-environment interaction
Investigator	Laboratory scientist	Epidemiologist	Interdisciplinary genetic epidemiologist
Mechanism	Metabolic activation	Metabolic activation, as well as: - metabolic deactivation - oncogene polymorphisms	All previous categories, plus: - DNA repair, - hormone and vitamin disposition, - cell cycle control, - neurotransmitter polymorphisms, - infection susceptibility, - disposition of lipids, - receptor polymorphisms
Limiting factors	Difficulty and inherent bias of phenotyping in field setting	Early genotyping often inadequate functional information	- multiple comparisons - ethical constraints
Advantages	Implicit is some information on function of measured trait	Avoids phenotyping and attendant sources of bias	- increasingly facile genotyping - non-invasive biospecimen collection - interdisciplinary approach - larger study size advantages of cohort design - studies exploit the human genome map, i.e. linkage disequilibrium mapping

The following are proposed as criteria for evaluating candidate genes for population studies.

1. Interindividual variation in the trait exists in the population

The historic rationale for considering hereditary differences in metabolic polymorphisms as being important in cancer etiology are the existence of substantial phenotypic differences between individuals (Lower *et al.*, 1979). These differences may be manifest as wide variation in urine or serum measurements of a metabolite (e.g. isoniazid, procainamide or sulfamethazine and the acetylation phenotype) (Drayer & Reidenberg, 1977), a 'metabolic ratio' (debrisoquine to 4-hydroxydebrisoquine in urine) (Gonzalez & Idle, 1994), or protein or mRNA directly detectable in liver or other relevant organ. The original meaning of a phenotype was a clinically recognizable state, but in the present context the focus is on 'biochemical' or 'molecular' phenotypes. Without important variation in a phenotype in the population, the rationale for examining a genotype would be weak. There are a great number of pharmacological factors (i.e. differences in absorption, distribution and excretion) and environmental factors (e.g. influence of age, gender, other drugs, physiological state, presence of disease) that can contribute to such variability. This variation must be eliminated to the greatest possible degree or by study design or controlled in the analysis.

Important interindividual variation is often manifest as clinical consequences following administration of the medications that are dependent on the particular pathway. For example, hereditary variation in hydralazine (NAT2) metabolism may explain susceptibility to systemic lupus erythematosus (Grant, 1993). Severe neurotoxicity following 5-fluorouracil administration is related to dihydropyrimidine dehydrogenase variation (Tuckman *et al.*, 1985). The pharmacogenetic literature contains many further examples (Kalow, 1962). Precise definition of the phenotype is vital because misclassification degrades the power to detect important differences. Selection of the proper measure to characterize it is also critical; for example, head size differs between phenylketonurics and normal individuals but this trait does not distinguish individuals with and without phenylketonuria as well as plasma phenylalanine measurements. Using the more precise laboratory study to define the phenotype permits a much clearer distinction of affected and non-affected individuals (Bogardus & Lillioja, 1992).

2. The gene is involved in a process related to carcinogenesis

While a complete inventory of all the processes that contribute to carcinogenesis remains to be elucidated, certain pathways unquestionably participate in neoplastic transformation. Early studies established that carcinogens require metabolic activation, and it was proposed that genetic control of activation (Ayesh *et al.*, 1984; Kellerman *et al.*, 1973) or elimination (Seidegard *et al.*, 1986; Lower *et al.*, 1979) might account for variation in tobacco-related cancer susceptibility. A broader appreciation of human carcinogenesis suggests other categories of genes that may control processes of equal relevance (Table 3). These include genes involved with DNA repair, chromosome stability, the activity of oncogene or tumour suppressor genes, cell cycle control or signal transduction, influence on hormonal or vitamin metabolism pathways, immune function, and neurotransmitter action.

The experience of the last few years provides ample evidence that our understanding of all the processes that influence carcinogenesis is incomplete. Thus apoptosis, telomerases, obesity, and addiction (behaviour) are all influenced by genetic variation and are likely to yield basic insights relevant to cancer during the next few years through investigations into the genetic determinants.

In addition, unusual genetic mechanisms have begun to be studied in the context of malignancy, including imprinting (Feinberg *et al.*, 1995), non-chromosomal inheritance, epigenetic mechanisms (e.g. methylation), and transgenerational effects (Tomatis, 1994).

The implications of genes that act at more than one biological hierarchical level need to be anticipated. For instance, if one gene acts to alter the disposition of carcinogens in tobacco smoke, and another (or the same one) acts to alter the likelihood that an individual will smoke, the analysis needs to take this into account. CYP2A6 and CYP2D6 are plausible candidates that have

possible roles in both nicotine metabolism and nitrosamine activation.

3. The trait exhibits an inheritance pattern consistent with Mendelian transmission

Variability in a trait may be observed in the general population and there may be a plausible reason to consider a relation to cancer. The next step is a demonstration that variation has a genetic basis. Family studies that demonstrate a pattern of inheritance consistent with Mendelian transmission are required to establish that a phenotype is genetic in origin, e.g. as demonstrated in kindreds with debrisoquine-deficient metabolizers (Evans *et al.*, 1980). It has become routine in many studies to confirm that transmission of purported variants exhibits a pattern consistent with Mendelian transmission by demonstrating it in commercially available DNA from multigeneration kindreds. In animals, controlled matings can be used to study the genetic transmission of traits. In human populations the proportion of offspring with the observed trait (or genotype) can be compared with that expected under a particular genetic hypothesis, an approach formally known as segregation analysis.

Supportive evidence for a genetic basis for variation can also be provided by twin studies, demonstrating that concordance of a particular trait is significantly greater between monozygotic than dizygotic twins. A demonstration that a particular trait exhibits a distribution in the population consistent with the predictions of Hardy-Weinberg equilibrium is another piece of supporting evidence.

Certain phenotypes such as "mutagen sensitivity" have exhibited strong associations with smoking-related cancers (Spitz & Hsu, 1994; Spitz *et al.*, 1995) but the precise nature of this 'host susceptibility' factor remains incompletely understood. Demonstrating that population variability is consistent with a genetic origin is no substitute for the demonstration of a clear pattern of inheritance in families. Lacking a genetic basis, phenotypes remain vulnerable to the classic criticisms when associations with diseases are reported, i.e. covariates linked with disease are responsible for phenotypic variation, or treatment or secondary metabolic consequences of the condition present in the host biases the phenotype measurement, resulting in misclassification. The latter problem is typical of case-control studies but is avoided in cohort studies if phenotyping is performed prior to the onset of disease.

4. Gene action exists in the relevant organ

Genes exhibit differential expression. The demonstration of gene expression in the organ or tissue of interest can lend plausibility to the hypothesis of a putative role in tumorigenesis. Alternatively, expression in the liver allows one to postulate that a carcinogen is activated in that organ, and that a carcinogenic product is eventually transported to the relevant organ. Thus, CYP1A1 is largely absent from the liver but present in the lung, lending some plausibility to disease associations focused on pulmonary pathology. CYP2D6 is expressed in the brain, giving rise to hypotheses suggesting an influence on behaviour. At least one recent report finds no evidence of expression in the lung (Kivisto *et al.*, 1997), arguing against a role in lung cancer. GSTM1 null genotype has been epidemiologically associated with lung cancer and exhibits at least some expression in the lung (Nakajima *et al.*, 1995; Anttila *et al.*,1993), although the highest pulmonary expression by a member of the GST family may be GSTP1. Various explanations have been offered for the weak but consistent finding of an association of GSTM1 null genotype with lung cancer, e.g. close linkage to a gene with similar function which is expressed in the lung (i.e. GSTM3), expression in the nasal mucosa (Gervasi *et al.*, 1991), or a more systematic role of the gene in preventing lipid peroxidation and secondary promotion, pulmonary localization not being critical. The epidemiological and mechanistic findings must ultimately be reconciled in order to provide a coherent explanation.

5. Gene location and characterization

The location and characterization of the gene contributes to our understanding in a number of ways. The structure of the gene may reveal similarities with other genes which suggest function or parallels with other organisms. The existence of pseudogenes may create difficulties in designing assays that are specific for the sequence of interest. For example, the nuclear enzyme poly (ADP-ribose) polymerase functions in DNA repair

and recombination. Bhatia *et al.* (1990) reported an association of a polymorphism of a related gene, probably a pseudogene, with various tumours. The finding may implicate the locus as a site of a nearby tumour suppressor gene rather than involvement of the specific gene itself in tumour susceptibility. The location of the mutation itself in relation to the gene locus also has significance. Most mutations of pathological significance occur in the coding sequence, with tandem repeats and CpG dinucleotides being favoured as particular hot spots. Mutations in intragenic noncoding and regulatory sequences may also result in altered function. The extent to which a mutation alters function may also depend on the precise effect on expression (reduced, abolished, augmented or ectopic), the degree to which the phenotype is expressed in the heterozygote, the proportion of cells in which the mutant gene is present (hereditary mutations are present in all cells, somatic changes usually only in a few), and occasionally the parental origin of the mutation (i.e. imprinting). (Miyagawa, 1998).

6. Polymorphisms and mutation

Detectable variation in a particular gene ranges from deletion of the entire gene, through missense and nonsense mutations that are highly likely to ablate function, to single base-pair changes that result in amino acid changes. Changes in base pairs at degenerate codons are the least likely to impact on function, along with minor intronic changes. Variable repeat polymorphisms in intronic (presumably regulatory) areas have been widely studied (examples are H ras vtr, the androgen receptor, and dopamine transporter protein).

The absence of a functional role for a genetic variant that has an otherwise convincing association with disease implies that an observed association is due to a) the observed variant being in linkage disequilibrium with some other truly functional gene, b) the variant having a non-obvious regulatory role, or c) chance.

7. Gene-gene interaction

In general, genes act in concert to exert effects. Genes often act in pathways, and therefore effects on one gene may induce compensatory changes in others. Critical pathways may have redundancies. This is the case, for example, for the p450 genes that exhibit overlapping substrate specificities. As mentioned earlier, genes may act at different hierarchical levels in the organism, i.e. gene (regulation of transcription), biochemical (activation of carcinogen), organ (specific toxicity) or organism (alteration of behaviour). Only a limited number of studies have examined the combined effects of two or more genes. The best example is the combined effect of CYP1A1 and GSTM1 in relation to lung cancer risk in Asians (Nakachi, 1993). The underlying complexity suggests that it is not surprising that studies of single genes and common cancers have not always yielded consistent findings. Considering multiple genes is more realistic biologically and should provide a more complete view of susceptibility in future studies in which there are both adequate numbers of subjects and technologies for multiple genotyping.

Certain biological effects are reasonably well established as dependent on one particular gene. For instance, CYP2D6 seems to exert complete control of the metabolic fate of a variety of medications. Consistent with this, metabolic consequences in deficient metabolizers are observed when these drugs are administered, i.e. side-effects at relatively low doses. However, even for model substrates of CYP2D6 such as dextromethorphan, some influence of other genes is likely (Ducharme *et al.*, 1996). Given that multiple genes with overlapping substrate specificities influence the fate of single distinct moieties, it seems highly likely that the complex mixture of carcinogens in tobacco smoke depends on an array of polymorphic genes for its ultimate effects. A good example is NNK, a nitrosamine carcinogen that may be activated by one of many p450s, including CYP2A6 and CYP2D6. The overlapping specificities of p450 enzymes are consistent with the observation that these enzymes vary commonly without obvious ill effects in the population. Since inactivating mutations are at least occasionally present for a number of p450s the particular gene cannot be essential for life; inactivating mutations would otherwise be under strong negative selection and redundant pathways would be even more common.

With regard to chronic diseases and, particularly, malignancy, it seems that we must invoke a complex of genes and exogenous exposure combinations to alter risk. Thus for the metabolic

polymorphisms, combinations of Phase 1 genes, combinations of Phase 2 genes, and Phase 1 and Phase 2 genes together may influence cancer risk in unexpected ways. To date, the best-studied genes with possible joint effects are CYP1A1 and GSTM1. There is at least some suggestion of both effect modification (Kihara *et al.*, 1995; Nakachi, 1993) in the population data as well as biological interaction (Vaury *et al.*, 1995).

8. Animal models

Animal work is increasingly providing important information relevant to genetic factors in humans and has probably been underutilized as a source of candidate genes. Animal models provide distinct advantages, among them the ability to control for numerous sources of variability, directly administer carcinogens, perform detailed genetic studies including controlled matings, obtain available tissue from all organs, and precisely fit numbers to power requirements. On the other hand, most animal work has focused on the 'familial' cancer genes because of their clear mechanistic relationship with cancer, rather than the subtler but more common effects of polymorphisms that influence metabolism. Closer working relationships and interdisciplinary training can clearly benefit the field, and epidemiological studies on highly plausible genes of interest which have inadequate functional information (e.g. CYP1A1 and CYP2E1) would be greatly facilitated by further animal work. Knockout mice will provide invaluable adjunct information to answer some of these questions (Nebert & Duffy,1997). In addition, new 'susceptibility' loci being reported from animal studies can be expected to serve as a fertile source of new ideas for human evaluation. Animal work can often provide important confirmation that a hypothesized susceptibility gene is a plausible candidate for human study. For example, the demonstration that gastrointestinal adenomas in mice with fully penetrant mutations of the APC gene were blocked by sulindac administration provides important support for the potential chemopreventive role of this medication, suggests that the mechanism involves decreased apoptosis in enterocytes, and encourages further investigation of this class of drugs in humans (Boolbol *et al.*,1996).

Animal work suggests that modifier loci can alter the effects of major cancer genes. An example is provided by the modifier locus for the APC gene, demonstrated in the multiple intestinal neoplasia (Min) mouse (Dietrich *et al.*, 1993). This work hints that susceptibility gene studies might profitably focus on the question of whether any of the susceptibility genes modify penetrance in individuals in kindreds with 'rare' mutations of cancer genes (e.g. Brca1 and breast cancer).

Murine cancer models parallel the human situation to a greater or lesser degree. For example, the NZB mouse exhibits a genetically regulated clonal proliferation of aneuploid cells resembling human chronic lymphocytic leukaemia. The CLL-like mice do not, however, exhibit involvement of the lymph nodes or liver as in the human disease (Philips *et al.*, 1992). Nevertheless, studying this model is likely to provide insights into the pathogenesis of the disorder as well as new candidate genes for human study, once the mouse locus is mapped.

It is also interesting to note that the genetics of certain tumours are quite well established in murine models. For example, genetic susceptibility to lung tumours has been well documented in mice for over a half century (Dragani *et al.*, 1995). A possible candidate based on homology with an implicated mouse locus is K-ras. K-ras mutations are frequently observed in human adenocarcinoma, although to date no specific heritable susceptibility locus has been demonstrated.

Finally, it appears that for tumours that have been well studied in the murine model, multiple susceptibility loci exist. This is certainly the case for lung (Dragani *et al.*, 1995) and has also been demonstrated for hepatocarcinogeneis in mice (Manenti *et al.*, 1994). In humans, where exposures are more varied, the genetics are more complex, and familial clustering is not apparent or rare (as is the case for most human tumours), it would seem plausible to suppose that susceptibility loci are also multiple.

9. Human studies: gene frequency

Investigators should know or determine the gene frequencies of the relevant alleles prior to formal association testing, because power calculations depend on this variable. There is a trade-off between the prevalence of the gene and the magnitude of risk that may be detected by a particular study size. The more common a variant, the

less likely it is to exhibit a strong association, but generally the more power there is to establish an association of a given magnitude. An example of this is the CYP2D6 gene that has a deficient genotype frequency of around 7% in European Caucasians, 4% in American Blacks, and 1% or less in Asians. Other factors being equal, there is moderate power to detect a given effect in Caucasians, weak power in an African-American population, and poor power in an Asian population.

10. Human studies: genotype-phenotype

The function of the genetic variants selected for study is of central importance. Unfortunately, functional information is often not available or is incomplete. Kellerman first studied the relationship of aryl hydrocarbon hydroxylase inducibility to lung cancer in 1974. This trait is dependent on CYP1A1, and polymorphisms in this gene have been described. Nevertheless, the precise relationship of the various polymorphisms to the phenotype remains incompletely understood. In the case of CYP1A1 there is some evidence that other genes in the same pathway such as the arnt receptor or Ah receptor play a role in influencing inducibility (Micka *et al.*, 1997). It is possible for measurement of the phenotype to be imprecise (measurement error) or non-specific (i.e. influenced by more than one gene), or for environmental factors to interfere (e.g. CYP1A2). For certain genes the proposed phenotype (i.e. CYP2E1 and chlorzoxazone) and genotype exhibit an irregular correspondence.

Methodological problems arise if the phenotype-genotype relationship is unknown. Particularly, in the absence of a facile measure of function the dominance relationship of the gene and directionality of the hypothesized association are unknown. Without this information, statistical power is weakened. For example, assuming a simple two-allele system, the risk group may be a genotype (AA, AB, or BB), presence/absence of the polymorphism (e.g. AA and AB vs. BB; implicit in this choice is some knowledge of the dominance relationship), or gene frequency (frequency of A) to examine in relation to disease. When multiple alleles are present these choices grow more complex and the danger of post-hoc assignment influencing findings becomes acute. It has been suggested that, in certain cases, heterozygotes may be the at-risk group, compared to both homozygotes. Such a possibility may be suggested by marked deviation from Hardy-Weinberg equilibrium conditions.

11. Human studies: relationship to disease

Ultimately, *in vitro* and animal findings must be confirmed in humans to establish their significance for human health and disease. It can be anticipated that interpretation of the vast genetic information derived from the Human Genome Project will require large well-designed studies in human populations. It is often poorly understood that genes that have been studied in families also require specific study in the general population (frequency permitting) to establish whether and to what degree they are associated with risk in that setting.

The point with regard to candidate gene selection is that findings from epidemiological studies comprise another category of evidence to focus both laboratory and population studies in the future on particular genes.

12. Ethnic variation

The polymorphic genes that involve metabolic polymorphisms almost universally exhibit important ethnic and racial variation. Ethnic variability is important for a number of reasons. First, gene frequency in the ethnic group targeted for study must be known in order to perform power calculations and estimate the size of the study needed. For example, a study of the relationship of CYP2D6-deficient metabolizers to a condition of interest would have reasonable power in a Caucasian population where 7% of subjects are poor metabolizers, but would be impossible in an Asian population where 1% or less are deficient metabolizers. Second, qualitatively different gene variants imply that different functional variants exist and possibly predominate in different ethnic groups. Studying a 'new' ethnic group with variants important in a different group may therefore miss the critical sources of variation. Third, population stratification (also known as ethnic admixture) is a potential source of bias. The classic example is the apparent association of the Gm 3,5,13,14 haplotype with non-insulin dependent diabetes in an admixed popu-

lation of Pima Indians. When the analysis was stratified according to the degree of Indian ancestry the association disappeared (Knowler *et al.*, 1988). Population stratification can be thought of as a confounding issue, and the problem should be greatly minimized by the usual measures for the control of confounding, including proper control selection and adjustment for ethnicity. Methods that use relatives of cases as controls can theoretically eliminate this problem (e.g. the transmission disequilibrium test (Spielman *et al.*, 1993)). However, the choice of relative controls poses difficulties in population-based studies. For example, enrolling parents of subjects with solid tumours, which typically have a median age of onset in the 60s, is typically not feasible. Fourth, the geographical variation that underlies ethnic variation probably holds important clues to the factors that drive this type of variation and may suggest determinants of disease susceptibility. The relationship of sickle cell disease/trait and malaria is the classic example.

Table 3. Categories of genes studied in relation to cancer susceptibility

Gene type	Examples	Cancer
Phase 1	CYP1A1 (Kellerman *et al.*, 1973)	Lung, others
	CYP1A2 (Lang *et al.*, 1994)	Colon, lung
	CYP2A6 (Gullsten *et al.*, 1997)	Liver
	CYP2C9 (London *et al.*, 1997)	Lung
	CYP2D6 (Ayesh *et al.*, 1984)	Lung, bladder
	CYP2E1 (Uematsu *et al.*, 1991, Hildesheim et al.,1997)	Lung, NPC
Phase 2	GSTM1 (Seidegard *et al.*, 1986)	Lung, bladder, others
	GSTT1 (Chen, H. *et al.*, 1996)	Myelodysplasia
	Epoxide hydrolase (Heckbert *et al.*, 1992)	Lung
	NAD(P)H quinone oxidoreductase (Schulz *et al.*, 1997)	Lung, kidney
	NAT1 (Bell *et al.*, 1995)	Colon, bladder
	NAT2 (Lower *et al.*, 1979; Cascorbi *et al.*, 1996)	Bladder, lung, others
Oncogene polymorphism	H-ras vtr (Krontiris *et al.*, 1985)	Lung, breast, others
Tumour suppressor gene	p53 (Murata *et al.*, 1996)	Various
	APC	Various
Nutrition	Vitamin D polymorphism (Taylor *et al.*, 1996)	Prostate
	Methylene-tetrahydrofolate reductase (Chen *et al.*, 1996)	Colon
	Ethanol (Harty *et al.*, 1997)	Oral cavity
DNA repair	Bleomycin sensitivity phenotype (Spitz & Hsu, 1994; Spitz *et al.*, 1995)	Lung, head and neck
Behaviour	DRD2 (Comings *et al.*, 1994)	No cancer studies to date
	Serotonin receptor (Lerman *et al.*, 1998)	No cancer studies to date
Inflammation	COX-2 (prostaglandin synthesis) (Boolbol *et al.*, 1996)	No cancer studies to date
Hormone	Testosterone (Maenpaa *et al.*, 1993)	Prostate, breast
	5 alpha reductase (Reichardt *et al.*, 1995)	Prostate
	Aromatase	Breast
	Androgen receptor (Irvine *et al.*, 1995)	Prostate
Immune	HLA	Lymphoproliferative (?)
Medications and xenobiotics	Chemotherapeutic agents	Various
	Dioxin	Various
Miscellaneous	Blood groups	Stomach
	Apolipoprotein	Stomach

Conclusions

A number of criteria for the selection of candidate genes have been offered and placed loosely in the historical context. None of them are absolute but none should be ignored. A promising epidemiological finding should stimulate laboratory efforts, while a new gene implicated in an animal model should be considered for study in the appropriate human condition. Findings that lack either the population or the laboratory component require cautious interpretation. It is likely that the complete picture of most malignancies will involve multiple genes and complex biological interactions.

A few predictions can be offered in the light of trends observed over the past years. Larger studies that include exposure information and incorporate numbers of genetic markers (increasingly available from work related to the Human Genome Project) are anticipated in the next few years. For the first time there should be opportunities to examine gene-gene and gene-environment interaction in an appropriate setting, i.e. of adequate power and with relevant exposure information.

The problem of multiple comparisons should be dealt with by sequentially performing studies in multiple populations. Even large settings can be expected to have a 'hypothesis generating' component. In spite of the proliferating genetic approaches, traditional epidemiological designs will remain the mainstay (i.e. association studies rather than family-based linkage/sib-pair designs) because power and exposure information are needed to detect the relatively modest effects of these genes.

References

Anttila, S., Hirvonen, A., Vainio, H., Husgafvel-Pursiainen, K., Hayes, J.D. & Ketterer, B. (1993) Immunohistochemical localization of glutathione S-transferases in human lung. *Cancer Res.*, 53, 5643-5648

Auerbach, A.D., Buchwald, M. & Joenje, H. Fanconi anemia. (1998) In: Vogelstein, B. & Kinzler, K.W., eds. *The genetic basis of human cancer.* New York, McGraw Hill, pp. 317-332

Ayesh, R., Idle, J.R., Richie, J.C., Crothers, M.J. & Hetzel, M.R. (1984) Metabolic oxidation phenotypes as markers for susceptibility to lung cancer. *Nature*, 312, 169-170

Barcellos, L.F., Klitz, W., Field, L.L., Tobias, R., Bowcock, A.M., Wilson, R., Nelson, M.P., Nagatomi, J. & Thomson, G. (1997) Association mapping of disease loci by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.*, 61, 734-747

Bhatia, K.G., Cherney, B.W., Hupppi, K., Magrath, I.T., Cossman, J., Sausville, E., Barriga, F., Johnson, B., Gause, B. & Bonney, G. (1990) A deletion linked to a poly(ADP-ribose) polymerase gene on chromosome 13q33-qter occurs frequently in a normal black population as well as in multiple tumour DNA. *Cancer Res.*, 50, 5406-5413

Bell, D.A., Stephans, E.A., Castranio, T., Umbach, D.M., Watson, M., Deakin, M., Elder, J., Hendrickse C., Duncan, H. & Strange, R.C. (1995) Polyadenylation polymorphism in the acetyltransferase 1 gene (NAT1) increases risk of colorectal cancer. *Cancer Res.*, 55, 3537

Bogardus, C. & Lillioja, S. (1992) Pima Indians as a model to study the genetics of NIDDM *J. Cell.Biochem.*, 48, 337-343

Boolbol, S.K., Dannenberg, A.J., Chadburn, A., Martucci, C., Guo, X.J., Ramonetti, J.T., Abreu-Goris, M., Newmark, H.L., Lipkin, M.L., DeCrosse, J.J. & Bertagnolli, M.M. (1996) Cyclooxygenase-2 overexpression and tumour formation are blocked by sulindac in a murine model of familial adenomatosis polyposis. *Cancer Res.*, 56, 2556-2560

Caporaso, N. & Goldstein, A. (1995) Cancer genes: single and susceptibility: exposing the difference. *Pharmacogenetics*, 5, 59-63

Cascorbi, I., Brockmöller, J., Mrozikiewicz, P.M., Bauer, S., Loddenkemper, R. & Roots, I. (1996) Homozygous rapid arylamine N-acetyltransferase (NAT2) genotype as susceptibility factor for lung cancer. *Cancer Res.*, 56, 3961-3966

Chen, H., Sandler D.P., Taylor, J.A., Shore, D.L., Liu, E., Bloomfield, C.D. & Bell D.A. (1996) Increased risk of myelodysplastic syndromes in individuals with glutathione theta 1 (GSTT1) gene defect. *Lancet*, 347, 295-297

Chen, J., Giovannucci, E., Kelsey, K., Rimm, E.B., Stampfer, M.J., Colditz, G.A., Spiegelman, D., Willett, W.C. & Hunter, D.J. (1996) A methylenetetrahydrofolate reductase polymorphism and the risk of colorectal cancer. *Cancer Res.*, 56, 4862-4864

Comings, D.E., Muhleman, D., Ahn, C., Gysin, R. & Flanagan, S.D. (1994) The dopamine D2 receptor gene: a risk factor for substance abuse. *Drug Alcohol Depend.*, 34, 175-180

d'Errico, A., Taioli, E., Xiang, C. & Vineis, P. (1996) Genetic metabolic polymorphisms and the risk of cancer: a review of the literature. *Biomarkers*, 1, 174-177

Dietrich, W.F., Lander, E.S., Smith, J.S., Moser, A.R., Gould, K.A., Luongo, C., Borenstein, N. & Dove, W. (1993) Genetic Identification of Mom-1, a major modifier locus affecting min-induced intestinal neoplasia in the mouse. *Cell*, 75, 631-639

Dragani, T.A., Manenti, G. & Pierotti, M.A. (1995) Genetics of murine lung tumors. *Advances Cancer Res.*, 67, 83-112

Drayer, D.E. & Reidenberg, M.M. (1977) Clinical consequences of polymorphic acetylation of basic drugs. *Clin.Pharmacol. Ther.*, 22(3), 251-258

Ducharme, J., Abdullah, S. & Wainer, I.W. (1966) Dextromethorphan as an *in vivo* probe for the simultaneous determination of CYP2D6 and CYP3A activity. *J. Chromatogr.*, 678, 113-128

Evans, D.A.P., Mahgoub, A., Sloan, T.P., Idle, J.R. & Smith, R.L. (1980) A family and population study of the genetic polymorphism of debrisoquine oxidation in a white British population. *J. Med. Genet.*,17, 102-105

Feinberg, A.P., Rainier, S. & DeBaun, M.R. (1995) Genomic imprinting, DNA methylation, and cancer. *J. Natl. Cancer Inst.*, 17, 21-26

Gervassi, P.G., Longo, V., Naldi, F., Panattoni, G. & Ursino, F. (1991) Xenobiotic-metabolizing enzymes in human respiratory nasal mucosa. *Biochem. Pharmacol.*, 41, 177-184

Gonzalez, F.J. & Idle, J.R. (1994) Pharmacogenetic phenotyping and genotyping. *Clin. Pharmacokinet.*, 26, 59-70

Grant, D.M. (1993) Molecular genetics of the N-acetyltransferases. *Pharmacogenetics*, 3, 45-50

Gullsten, H., Agundez, J.A.G., Benitez, J., Laara, E., Ladero, J.M., Diaz-Rubio, M., Fernandez-Salguero, P., Gonzalez, F., Rautio, A., Pelkonen, O. & Raunio, H. (1997) CYP2A6 gene polymorphism and risk of liver cancer and cirrhosis. *Pharmacogenetics*, 7, 247-250

Harty, L.C., Caporaso, N.E., Hayes, R.B., Winn, D.M., Bravo-Otero, E., Blot, W.J., Brown, L.M., Armenian, H.K., Fraumeni, J.F.F. & Shields, P.G. (1997) Alcohol dehydrogenase 3 genotype and risk of oral and pharyngeal cancers. *J. Natl. Cancer Inst.*, 89, 1698-1705

Heckbert, S.R., Weiss, N.S., Hornung, S.K., Eaton, D.L. & Motulsky, A.G. (1992) Glutathione S transferase and epoxide hydrolase activity in human leukocytes in relation to risk of lung cancer and other smoking-related cancers. *J. Natl. Cancer Inst.*, 84, 414-421

Hildesheim, A., Anderson, L.M., Chen, C.-J., Brinton, L.A., Daly, A.K., Reed, C.D., Chen, I.-H., Caporaso. N.E., Hsu, M.-M., Chen, J.-Y., Idle, J.R., Hoover, R.N., Yang, C.-S. & Chhabra, S.K. (1997) CYP2E1 genetic polymorphism and the risk of nasopharyngeal carcinoma in Taiwan. *J. Natl. Cancer Inst.* , 89, 1207-1212

Irvine, R.A., Yu, M.C., Ross, R.K. & Coetzee, G.A. (1995) The CAG and GGC microsatellites of the androgen receptor gene are in linkage disequilibrium in men with prostate cancer. *Cancer Res.*, 55, 1937-1940

Kalow, W. (1962) *Pharmacogenetics: heredity and response to drugs.* Philadelphia, W.B. Saunders

Kellerman, G., Shaw, C.R. & Luyten-Kellerman, M. (1973) Aryl hydrocarbon hydroxylase inducibility and bronchogenic carcinoma. *New Engl. J. Med.*, 289, 934-937

Kihara, M., Kihara, M. & Noda, K. (1995) Risk of smoking for squamous and small cell carcinomas of the lung modulated by combinations of CYP1A1 and GSTM1 gene polymorphisms in a Japanese population. *Carcinogenesis*, 16, 2331-2336

Kivisto, K.T., Griese, E.-U., Stuven, T., Fritz, P., Friedel, G., Kroemer, H.K. & Zanger, U.M. (1997) Analysis of CYP2D6 expression in human lung: implications for the association between CYP2D6 activity and susceptibility to lung cancer. *Pharmacogenetics*, 7, 295-302

Knowler, W.C., Williams, R.C., Pettitt, D.J. & Steinberg, A.G. (1988) Gm 3;5;13;14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am. J. Hum. Genet.*, 43, 520-526

Krontiris, T.G., DiMartino, N.A., Colb, M. & Parkinson, D.R. (1985) Unique allelic restriction fragments of the human Ha-ras locus in leukocyte and tumour DNAs of cancer patients. *Nature*, 313, 369-374

Lander, E. & Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.*, 11, 241-247

Lang, N.P., Butler, M.A., Massengill, J., Lawson, M., Stotta, R.C., Haver-Jensen, M. & Kadlubar, F.F. (1994) Rapid metabolic phenotypes for acetyltransferase and cytochrome P4501A2 and putative exposure to food-borne heterocyclic amines increases risk for colorectal cancer or polyps. *Cancer Epidemiol. Biomarkers Prev.*, 3, 675-682

Lerman, C., Shields, P.G., Audrain, J., Main, D., Cobb, B., Boyd, N.R. & Caporaso, N. (1998) The role of the serotonin transporter gene in cigarette smoking. *Cancer Epidemiol. Biomarkers Prev.*, 7, 253-255

London, S.J., Sullivan-Klose, T., Daly, A.K. & Idle, J.R. (1997) Lung cancer risk in relation to the CYP2C9 genetic polymorphism among Caucasians in Los Angeles County. *Pharmacogenetics*, 7, 401-405

Lower, G.M., Nilsson, T., Nelson, C.E., Wolf, H., Gamsky, T.E. & Bryan, G.T. (1979) N-acetyltransferase phenotype and risk in urinary bladder cancer: approaches in molecular epidemiology. Preliminary results in Sweden and Denmark. *Environ. Health Perspect.*, 29, 71-79

Maenpaa, J., Pelkonen, O., Cresteil, T. & Rane, A. (1993) The role of cytochrome p450 3A (CYP3A) isoforms in oxidative metabolism of testosterone and benzphetamine in human adult and fetal liver. *J. Steroid Biochem.*, 44, 61-67

Manenti, G., Binelli, G., Gariboldi, M., Canzian, F., DeGregorio, L., Falvella, F.S., Dragani, T.A. & Pierotti, M.A. (1994) Multiple loci affect genetic predisposition to hepatocarcinogenesis in mice. *Genomics*, 23, 118-124

McGlynn, K.A., Rosvold, E.A., Lustbader, E.D., Hu, Y., Clapper, M.L., Zhou, T., Wild, C.P., Xia, X.L., Baffoe-Bonnie, A., Ofori-Adjei, D., Chen, G.C., London, W.T., Shen, F.M. & Buetow, K.H. (1995) 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

Micka, J., Milatovich, A., Menon, A., Grabowski, G.A., Puga, A. & Nebert, D.W. (1997) Human Ah receptor (AHR) gene: localization to 7p15 and suggestive correlation of polymorphism with CYP1A1 inducibility. *Pharmacogenetics*, 7, 95-102

Miyagawa, K. Genetic instability and cancer. (1998) *Int. J. Hematol.*, 67 (1), 3-14

Murata, M., Tagawa, M., Kimura, M., Kimura, H., Watanabe, S. & Saisho, H. (1996) Analysis of a germ line polymorphism of the p53 gene in lung cancer patients; discrete results with smoking history. *Carcinogenesis*, 17, 261-264

Nakachi, K., Imai, K., Hayashi, S. & Kawajiri, K. (1993) Polymorphism of the CYP1A1 and glutathione-S-transferase genes associated with susceptibility to lung cancer in relation to cigarette dose in a Japanese population. *Cancer Res.*, 53, 2994-2999

Nakajima, T., Elovaara, E., Anttila, S., Hirvonen, A., Camus, A.-M., Hayes, J.D., Ketterer, B. & Vainio, H. (1995) Expression and polymorphism of glutathione S-transferase in human lungs: risk factors in smoking-related lung cancer. *Carcinogenesis*, 16, 707-711

Nebert, D.W. & Duffy, J.J. (1997) How Knockout mice will be used to study the role of drug-metabolizing enzymes and their receptors during reproduction and development, and in environmental toxicity, cancer and oxidative stress. *Biochem. Pharmacol.*, 53, 249-254

Nowell, P.C. & Hungerford, D.A. (1960) A minute chromosome in human chronic granulocytic leukemia. *Science*, 132, 1497

Oda, Y., Tanaka, M. & Nakanishi, I. (1994) Relation between the occurrence of K-ras gene point mutations and genotypes of polymorphic N-acetyltransferase in human colorectal cancer carcinomas. *Carcinogenesis*, 15(7), 1365-1369

Paul, J.R., Melnick, J.L. & Riordan JT (1952) Comparative neutralizing antibody patterns to Lansing (type 2) poliomyelitis virus in different populations. *Am. J. Hygiene*, 56, 232-251

Pensotti, V., Radice, P., Presciuttini, S., Calistri, D., Gazzoli, I., Perez, A.P.G., Mondini, P., Buonsanti, G., Sala, P., Rossetti, C., Ranzani, G.N., Bertario, L. & Pierotti, M.A. (1997) Mean age of tumour in hereditary nonpolyposis colorectal cancer (HNPCC) families correlates with the presence of mutations in DNA mismatch repair genes. *Genes, Chromosomes and Cancer*, 19, 135-142

Phillips, J.A., Mehta, K., Fernandez, C. & Raveche, E.S. (1992) The NZB mouse as a model for chronic lymphocytic leukaemia. *Cancer Res.*, 52, 437-443

Reichardt, J.K.V., Makridakis, N., Henderson, B.E., Yu, M.C., Pike, M.C. & Ross, RK. (1995) Genetic variability of the human SRD5A2 gene: implications for prostate cancer risk. *Cancer Res.*, 55, 3973-3975

Risch, N. & Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, 273, 1516-1517

Scrivner, C.R., Beaudet, A., Valle, D. & Sly, W.S., eds. (1995) *The metabolic and molecular basis of inherited disease. Volumes I-III, 7th Ed.* New York, McGraw-Hill

Seidegard, J., Pero, R.W., Miller, D. & Beattie, E.J. (1986) A glutathione transferase in human leukocytes as a marker for susceptibility to lung cancer. *Carcinogenesis*, 7, 751-753

Schulz, W.A., Krummeck, A., Rosinger, I., Eickelmann, P., Neuhaus, C., Ebert, T., Schmitz-Drager, B.J. & Sies, H. (1997) Increased frequency of a null-allele for NAD(P)H: quinone oxidoreductase in patients with urological malignancies. *Pharmacogenetics*, 7, 235-239

Soini, Y., Chia, S.C., Bennett, W.P., Groopman, J.D., Wang, J.S., De Benedetti, V.M., Cawley, H., Welsh, J.A., Hansen, C., Bergasa, N.V., Jones, E.A., Di Bisceglie, A.M., Trivers, G.E., Sandoval, C.A., Calderon, I.E., Munoz Espinosa.L.E. & Harris, C.C. (1996) An aflatoxin-associated mutational hotspot at codon 249 in the p53 tumor-suppresser gene in HCC. *Cancer*, 17(5), 1007-1020

Spielman, R.S., McGinnis, R.E. & Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, 32, 506-512

Spitz, M.R. & Hsu, T.C. (1994) Mutagen sensitivity as a marker of cancer risk. *Cancer Detect. Prev.*, 18, 299-303

Spitz, M.R., Hsu, T.C., Wu, X., Fueger, J.J., Amos, C.I. & Roth, J.A. (1995) Mutagen sensitivity as a biological marker of lung cancer risk in African Americans. *Cancer Epidemiol. Biomarkers Prev.*, 4, 99-103

Taylor, J.A., Hirvonen, A., Watson, M., Pittman, G., Mohler, J.L. & Bell, D.A. (1996) Association of prostate cancer with vitamin D receptor gene polymorphism. *Cancer Res.*, 56, 4108-4110

Tomatis, L. (1994) Transgenerational arcinogenesis: a review of the experimental and epidemiological literature. *Jpn. J. Cancer Res.*, 85, 443-454

Tuchman, M., Stoeckeler, J.S., Kiang, D.T., O'Dea, R.F., Ramnraine, M.L. & Mirkin, B.L. (1985) Familial pyrimidemia and pyrimidinuria associated with severe fluorouracil toxicity. *New Engl. J. Med.*, 313, 245-249

Uematsu, F., Kikuchi, H., Motomiya, M., Abe, T., Sagami, I., Ohmachi, T., Wakui, A., Kanamara, R. & Watanabe, M. (1991) Association between restriction fragment length polymorphism of human cytochrome P450IIE1 gene and susceptibility to lung cancer. *Jpn. J. Cancer Res.*, 82, 254-256

Wistuba, I.I., Lam, S., Behrens, C., Virmani, A.K., Fong, K.M., LeRiche, J., Samet, J.M., Srivastava, S., Minna, J.D. & Gazdar, A.F. (1997) Molecular damage in the bronchial epithelium of current and former smokers. *J. Natl. Cancer Inst.*, 89, 1366-1372

Vaury, C., Laine, R., Noguiez, P., de Coppet, P., Jaulin, C., Praz, F., Pompon, D. & Amor-Gueret, M. (1995) Human glutathione S- transferase M1 null genotype is associated with high inducibility of cytochrome p450 1A1 gene transcription. *Cancer Res.*, 55, 5520-5523

Neil Caporaso
Pharmacogenetic Section,
Genetic Epidemiology Branch,
National Cancer Institute,
EPN 439,
Rockville, MD 20892, USA