OPINION

The balance between heritable and environmental aetiology of human disease

Kari Hemminki, Justo Lorenzo Bermejo and Asta Försti

Abstract | The Human Genome Project and the ensuing International HapMap Project were largely motivated by human health issues. But the distance from a DNA sequence variation to a novel disease gene is considerable; for complex diseases, closing this gap hinges on the premise that they arise mainly from heritable causes. Using cancer as an example of complex disease, we examine the scientific evidence for the hypothesis that human diseases result from interactions between genetic variants and the environment.

The present flood of genetic and genomic data and references to genomic medicine might give the impression that "…most diseases are the result of the interactions of multiple genes and environmental factors…"[1] or that "…almost all human diseases result from interactions between genetic variants and the environment."[2] However, these statements do not seem to reflect that many studies point to a predominantly environmental causation of complex diseases[2–8], or the limited progress that is widely acknowledged in the genetic analysis of common diseases[6,9–11]. The main question that remains in human disease genetics is that of aetiology: how much do we understand about the heritable and environmental causation?

Heritable causes of complex diseases remain largely elusive, despite tremendous efforts to understand them. Partly, this is because the genes that underlie complex diseases are thought to have weak effects on disease susceptibility, conveying familial clustering with complex non-Mendelian patterns, which explains the connotation 'complex disease'. By contrast, the genes that are identified often confer a high risk of disease (high penetrance) to carriers[12,13]. Technological excellence in genomics does not automatically lead to benefits in human health, which could require a true understanding of the aetiology of these 'complex' or 'multifactorial' diseases, and which, we argue, require a true understanding of the role of the environment[9,14].

Here we explore the magnitude of complex-disease heritability and the role of the environment in their aetiology. The basic dilemma in complex-disease genomics in the developed countries is that heritable causes will be difficult to find, because environmental factors have increased the background incidence to over 10 times the level that is found in the developing countries[4,6,15]. Moreover, the inferred gene–environment interactions, that is, the expression of the heritable factors against a high background of environmentally dominated disease, are poorly understood.

The quest for aetiological understanding is shared by all complex diseases[16]. Here we use cancer as an example of a complex disease to examine the heritable and environmental aetiology. Our choice is motivated by the existence of reasonably uniform diagnostics, and the availability of global incidence figures and a wealth of aetiological and mechanistic data.

We use 'genetic' as defined in the Dorland's Illustrated Medical Dictionary, meaning "…pertaining to or determined by genes", and 'hereditary' and 'heritable' meaning "…genetically transmitted from parent to offspring." Note that 'genetic' does not differentiate between the germ line and the somatic origin, in contrast to 'hereditary' and 'heritable'. 'Heritability' is the phenotypic variance that is attributable to genetic effects[17]. Heritability specifically considers variation in the occurrence of the disease or trait between individuals.

So, gene *X* might cause disease Y, but if there is no variation in gene *X* in the population then it would not contribute to the heritability of disease Y, nor would it cause familial clustering of the disease. It would also be almost impossible to prove that gene *X* causes disease Y[18]. It would be equally impossible to prove environmental origins of a disease if there was no variation in environmental exposures in the populations. In this context, 'environment' is any influencing factor that is not inherited, including sporadic, random causation of disease.

Measures of heritable effects

Familial clustering of a disease is a direct indicator of a possible heritable cause, provided that environmental sharing can be excluded[5,19] (BOX 1). If familial risk is lacking, the likelihood of a heritable influence is also small[17]. Because genes are inherited from parents, any gene that shows an association with a disease also contributes to its familial risk. Association studies typically measure the frequencies of variant (mutant) genotypes in series of cases and controls, and assess the differences statistically by genotype odds ratios. There are no generally accepted definitions of 'high' and 'low' risk genotypes; for rare disease genes (allele frequencies <0.01), a risk of over 10.0 might be 'high' and a risk of below 2.0 might be 'low', but for common variants (allele frequencies >0.1), a risk of over 3.0 might be 'high' and a risk of below 1.5 might be 'low'.

Several simulations of the interdependence of the genotype odds ratio and the familial risk have been published[20–23]. TABLE 1 shows the dependence of familial risk (parent–offspring relationship) on the allele frequency of the susceptibility gene, and on the genotype odds ratio for a dominant mode of inheritance (for a more detailed discussion, see REFS 24,25). Our calculations in TABLE 1 show that rare alleles influence familial risks to a limited extent, even if their genotype odds ratios are high, which is similar to what occurs for common variants with low risk. The data show that a genotype odds ratio of 2.0 only marginally increases familial risk (up to 1.06). Even a genotype odds ratio of 5.0 increases the familial risk to no higher than 1.38. To explain a familial risk of 1.8 for breast cancer (which is characteristic of most types of cancer[26]), at least five genes with similar effects to the breast cancer gene *BRCA1* would be required; the mutant allele frequency of *BRCA1* is about 0.0005 (REF. 27). If the familial risk was entirely due to low-penetrance genes, the number of genes would need to be much higher.

This example shows that common susceptibility variants with low genotype odds ratios only marginally influence the familial risk. They might, nevertheless, have large

Box 1 | **Epidemiological and genetic epidemiological terms**

Epidemiological methods are used to measure the occurrence of a disease in the population and to identify factors that affect variation in disease patterns. Epidemiology has traditionally focused on environmental factors, with the aim of understanding environmental causes of the disease. The related field, genetic epidemiology, aims to understand the role of heritable factors in disease causation. Some of the terms and measures are identical, with the distinction that the term 'exposure' describes an environmental factor in epidemiology, whereas in genetic epidemiology, it descibes a gene. Increasingly, the two fields are interacting in population recruitment, genomic technologies and statistical analysis of complex diseases, which could eventually result in a unified understanding of disease aetiology.

Incidence
The occurrence of new cases of disease in a population over a specified time period; for cancer, the annual incidence is usually quoted as the number of cases in 100,000 people (or people each year).

Prevalence
The total number of disease cases (old and new) in the population.

Odds ratio
The risk of disease in exposed individuals compared with unexposed individuals, as used in case–control studies. For rare diseases, the odds ratio is a close estimator of the relative risk.

Relative risk
The risk of disease in exposed individuals compared with unexposed individuals, as used in follow-up (cohort) studies. For example, the relative risk of lung cancer in active smokers is about 20; the relative risk of lung cancer in non-smokers who are married to a smoker is 1.2–1.3 (REF. 30).

Standardized incidence ratio
A relative risk measure that is adjusted for variables such as age.

Population attributable fraction
The proportion of cases of a disease in a population that are explained by an exposure or genotype. As such, this fraction of the number of cases of disease would disappear if the exposure or the genotype did not exist.

Significant risk
The observed risk that is not likely to be explained by chance. Statistical significance is commonly defined through 5% or 1% confidence intervals or *P*-values, which allow a 5% or 1% chance occurrence, respectively. Statistical significance depends on the magnitude of the risk and the sample size. In large studies, small odds ratios such as 1.1 might be statistically significant but biologically meaningless. Such results could be due to unobserved variables (confounders).

Genotype odds ratio
An odds ratio that refers to those who have a certain genotype compared with those who have another genotype. It is sometimes referred to as 'genotype relative risk'.

Familial risk
The risk in those whose relatives (probands) have a particular disease compared with the risk in those whose relatives lack the disease. It can be defined through a specified relationship, such as parent–offspring, siblings or first-degree relatives. The impact of familial risk for the population and for the individual is higher for common diseases compared with rare diseases. It is sometimes referred to as 'recurrence risk'.

Individual risk
The risk for an individual to contract a disease by a defined age. The individual risk is used in clinical genetic counselling. The individual risk can be high in heritable diseases of high penetrance, even though the diseases are rare.

Individualized medicine
A novel, largely futuristic area of genomic medicine, in which an individual's genetic make up is used to predict his or her risk profile.

Twin study
A classical method of inferring the heritability of a trait through the correlation of that trait in pairs of monozygotic (genetically identical) and dizygotic (on average, 50% genetically identical) twins. The twin model assumes that monozygotic and dizygotic twins share environmental exposures to an equal extent. The twin model cannot resolve gene–environment interactions.

Table 1 | **Effect of genotype odds ratio and allele frequency on familial risk**

Allele frequency	Genotype odds ratio				
	1.5	2.0	3.0	5.0	10.0
0.01	1.00	1.01	1.04	1.13	1.57
0.05	1.01	1.04	1.12	1.36	1.99
0.1	1.02	1.05	1.15	1.38	1.80
0.2	1.02	1.06	1.14	1.28	1.46
0.5	1.01	1.02	1.04	1.06	1.08

population attributable fractions of cases of the disease that are caused by the variant, implying that the absence of these variants (for example, in a population that lacks disease alleles) would prevent a large proportion of disease cases. When a candidate-gene study targets a common variant, the *a priori* expectation is that the odds ratio is low and the population attributable fraction is high. Examples of the dependence of the population attributable fraction on allele frequency and genotype odds ratio are shown in TABLE 2, on the basis of a dominant model. Even a genotype odds ratio of 1.5 causes a high population attributable fraction at high allele frequencies, for example, being 27.3% with an allele frequency of 0.5. This allele frequency and a genotype odds ratio of 3.0 would explain over half of the disease occurrence. For comparison, mutations in *BRCA1*, which has a high-penetrance, show a population attributable fraction of 1%. Consequently, small errors in estimates of risk for common genes might have larger effects on population attributable fractions than the entire effects of known high-penetrance genes.

The population attributable fraction of disease-susceptibility genes cannot exceed 100% and, as we show below, the scientific justification for heritable causation does not extend beyond familial and twin data. For non-Mendelian conditions, the familial risk has a limited predictive power, but the existence of heritability beyond familial clustering is based on arbitrary assumptions. The population attributable fraction and the familial risk that are associated with particular genes have to be in concordance: if all the disease-susceptibility genes were identified, their effects should completely explain the familial risks (if environmental causes for familial clustering are excluded; see below for an example).

Genes and environment

The discussion of the aetiology of human disease originates from the dichotomy of nature and nurture in twin studies, well before the discovery of the double helix[28]. Since the 1960s, some 75–90% of cancer has been thought to be environmental (that is, not heritable)[3,29]. In epidemiological terms, the population attributable fraction of environmental factors is considered to be up to 90% for all types of cancer. For lung cancer, the predictions have been shown to be correct[30]. For coronary heart disease, stroke and type 2 diabetes, population attributable fractions of known environmental factors are also thought to be over 70% (REF. 4).

More recently, the epidemiologically founded description of human cancer causation has become more complicated. First, it has been shown that some environmental factors, such as tobacco smoking and asbestos exposure, interact and jointly create a higher risk of lung cancer than the sum of the separate risks[31]. Second, the effects of environmental exposures might be transmitted at the cellular level through mechanisms that could vary between individuals. A uniform exposure would elicit different effects depending on an individual's genetic make up. Although relatively little is known about the exact carcinogenic mechanisms of environmental insults, many known carcinogens could potentially elicit heritable effects at many levels, including carcinogen metabolism, DNA repair, cell-cycle control and apoptosis[32].

Theoretical considerations have led some authors to claim that both heritable and environmental factors each cause 100% of disease[2,33]. In a widely cited statement, Rothman and Greenland argued that: "If all genetic factors that determine disease are taken into account, whether or not they vary within populations, then 100% of disease can be said to be inherited. Analogously, 100% of any disease is environmentally caused…"[33] As read, the message states that genes cause all diseases, which, although true, is as useful as stating that proteins cause all diseases or that cells cause all diseases. Unfortunately, because of a semantic confusion (between the meaning of genetic and heritable), this truism is often misinterpreted as implying that the heritable and environmental causes of all diseases each add up to 100%. If there is no allelic variation in the gene in the population, then it does not contribute to the heritability of the disease, as discussed above.

In a single population in which a population attributable fraction is measured for heritable and environmental factors, and for any of their interactions, the result cannot exceed 100% of the disease. A favourite example in this context has been phenylketonuria[33] — an inborn error of metabolism in which the patients are unable to metabolize phenylalanine; the mental retardation that is a possible outcome of this disorder can be completely prevented by a diet that is low in phenylalanine. It is a rare example of a complete disease causation by a gene–environment interaction. It is important to consider that a limited knowledge of the disease aetiology might lead to the incorrect belief that the disease is 100% heritable (when only genetic factors are considered) or 100% environmental (when only dietary factors are considered).

Heritability in cancer

Inherited cancer syndromes of high penetrance with no appreciable environmental influence are thought to account for 1–2% of cancers, at most[34]. Low penetrant familial cancer could amount to 10% of cancers, but this proportion depends, for example, on whether only first-degree or more distant

Table 2 | **Effect of disease variant on population attributable fraction**

Allele frequency	Genotype odds ratio				
	1.5	2.0	3.0	5.0	10.0
0.01	1.0	2.0	3.8	7.4	15.2
0.05	4.6	8.9	16.3	28.1	46.7
0.1	8.7	16.0	27.5	43.2	63.1
0.2	15.3	26.5	41.9	59.0	76.4
0.5	27.3	42.9	60.0	75.0	87.1

relatives are included. Familial risks for most cancers are around 2.0, and familial population attributable fractions among first-degree relatives range from 9.1% for prostate cancer to 0.2% for connective-tissue tumours[35].

Twin studies, which measure the concordance of a disease in monozygotic and dizygotic twins, have been the classical way to examine disease aetiology, even though the results can be difficult to interpret because of the unquantifiable gene–environment interactions[17]. If such interactions were higher than additive, they would erroneously increase the heritability estimates. According to a Nordic twin study, the heritability estimates that were derived for colorectal (35%), breast (27%) and prostate (42%) cancers were the only significant estimates for site-specific cancers[36]. The non-shared, random environment was the main contributor to all cancers, which is in line with other evidence on the importance of environmental effects in cancer. Concordance rates for monozygotic twin pairs by age 75 years were only 11% for colorectal cancer, 13% for breast cancer and 18% for prostate cancer, and these were lower in dizygotic twins (5%, 9% and 3%, respectively)[36]. Monozygotic twins share 100% of their genes and much of their environmental experiences, particularly early in life, more so than any other pair of human beings. Therefore, the low concordance rates agree with other data on cancer incidence and provide little support for strong heritable effects in cancer.

A challenging mental exercise would be to compare population attributable fractions for the environmental and heritable causes of cancer, thereby advancing aetiological understanding. If the western population was to live in the same conditions as the populations of developing countries, the risk of cancer would decrease by 90%, provided that viral infections and mycotoxin exposures could be avoided[37] (see below). Eradication of hereditary cancer syndromes would reduce the cancer burden by 1%, and up to 10% of the population would be saved if all familial cancers could be avoided. In some cases, however, familiar clustering can be explained by environmental factors. In Iceland, where familial clustering of cancer has been observed over many generations[38,39], the risks for spouses exceeded those for second-degree relatives in lung, stomach and colon cancers, implying that environmental effects contribute to the familial clustering of these cancers in this population. It is not possible to show a consensus for the inclusion of other heritable factors, because there are few replicated findings on low-risk genes that predispose to cancer. Importantly, however, we must acknowledge an almost complete ignorance of the relevant gene–environment interactions — as data accumulate, causes that now seem to be environmental could turn out to be gene–environment interactions (as in the phenylketonuria example above).

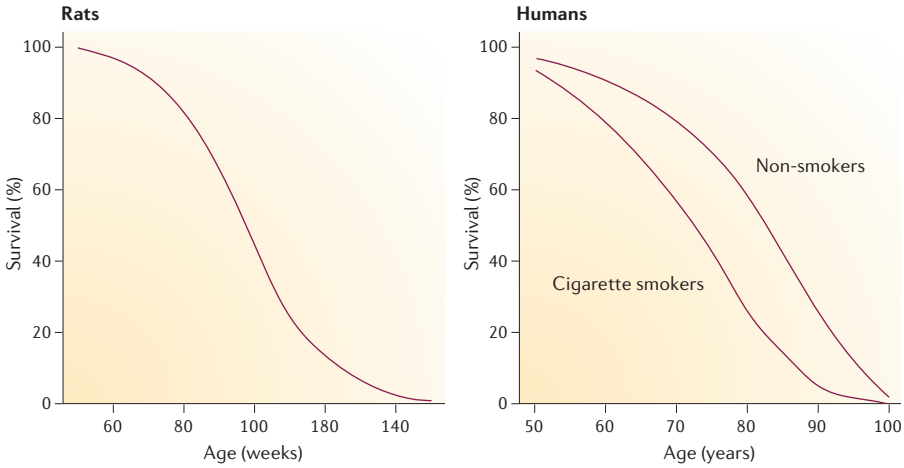

Figure 1 | **Survival in rats and humans — inter-individual variation.** Survival of the control inbred Sprague–Dawley female rats (N=150) in the aspartame bioassay is shown on the left[40]. Survival of 34,439 British male smoking and non-smoking doctors, whose cause-specific mortality was followed for 50 years, is shown on the right[40]; the curves show survival when follow up was started at the age of 35 years (100% survival). The strikingly similar shapes of the curves indicate that inter-individual differences that are observed for genetically identical rats housed in a standard environment are due to random stochastic processes. Figure modified with permission from REF. 40 and REF. 41 © (2004) BMJ Publishing Group Ltd.

The argument that not all smokers develop lung cancer is commonly used in favour of the importance of genetic factors. However, variation between individuals is an inherent property of all biological processes, as shown by the comparison of survival curves for inbred experimental animals and outbred humans in FIG. 1. The left panel shows results from a lifetime bioassay of inbred rats that were housed in a controlled environment[40]. The right panel shows the survival of British doctors in the Doll and Peto study of the effects of smoking[41]. In both panels, the *x* axis covers lifespan from almost 100% to nearly 0% survival. The similarity in the shapes of the three survival curves is striking, indicating that random effects cause a survival variability in genetically identical rats housed in standard conditions that is approximately equal to the survival variability in smoking and non-smoking doctors. Stochastic variation influences the fate of smokers, just as it influences any disease risk, providing no clues about genetic causes.

Environmental origins of cancer

The International Agency for Research on Cancer (IARC) has collected quality-assured incidence data on various types of cancer from around the world[42], which were used as the primary evidence for the environmental origin of cancer[3,29]. The highest and lowest reported age-standardized incidences for four of the main neoplasms — colon cancer, breast cancer, prostate cancer and non-Hodgkin lymphoma — are shown in TABLE 3. The differences in incidence range from 200-fold for prostate cancer to 13-fold for female non-Hodgkin lymphoma. Although the extreme values could be caused by small random variations, the highest incidence rates are well representative of the level that is generally observed in the developed countries. Analogously, the lowest rates closely represent the rates for the large Asian and African populations. These four types of cancer were selected because they are common in the developed countries and they share few risk factors, except for age. A similar comparison for any type of cancer would show at least a 10-fold difference between the regions of low and high incidence[42]. Some cancers, such as those of the liver, oesophagus and cervix, are more common in the developing countries than in the developed countries. Liver cancer is associated with the hepatitis B and C viruses and with ingestion of mycotoxin aflatoxin B1, whereas cervical cancer is associated with human papilloma virus infection[37].

Table 3 | **The highest and lowest age-adjusted cancer incidence in 100,000 people**[42]

Gender	Cancer incidence and geographical location			
	Highest		*Lowest*	
Colon cancer				
Men	59.2	Japan, Hiroshima	0.5	The Gambia
Women	28.0	Japan, Hiroshima	1.0	India, Karunagappally
Breast cancer				
Women	114.9	Uruguay	7.0	The Gambia
Prostate cancer				
Men	202.0	USA, Detroit (black)	1.1	China, Qidong
Non-Hodgkin lymphoma				
Men	24.6	USA, San Francisco (white)	2.0	Mali
Women	13.1	Belgium, Flanders	1.0	The Gambia

Many published migrant studies have shown that cancer incidence changes on migration, pointing to a predominant environmental contribution to cancer causation[3,43–46]. Moreover, there have been strong incidence changes in single regions. For example, during the operation of the Swedish Cancer Registry, from 1958 to 2003, the incidence of male melanoma increased 7.7-fold, squamous-cell skin cancer increased 4.1-fold, prostate cancer and non-Hodgkin lymphoma both increased 3.2-fold, and breast cancer increased 2.2-fold[47]. At the same time, the incidence of male gastric cancer decreased 3.4-fold. Such changes can also be found in other registration systems with long periods of follow up, such as the Connecticut Tumor Registry[48]. In Japan, which is historically a low-risk area for colon cancer, there has been a dramatic increase in the incidence of this disease (some 10-fold in men between 1960 and 1990 according to the Miyagi Cancer Registry[49]). According to TABLE 3, the highest male and female rates for colon cancer are scored in Hiroshima.

The driving forces for the changes in cancer incidence that are discussed above are clearly environmental, but their cellular effects could be transmitted through gene products as a result of gene–environment interactions. These observations should help us to qualify some features of the underlying effects and mechanisms. First, environmental factors must be diverse and widespread, such as overall energy intake, in order to affect practically all cancer types that are not known to share risk factors. Second, by the same logic, the genes that are assumed to be responsive to these environmental factors must also be diverse. These genes probably constitute the vast set of genes that are mutated in sporadic cancers[50]. Many of these genes have important cellular functions (such as P53), and therefore show low functionally relevant allelic variation in the population. Because of their large number and limited allelic variation, they would not be detectable by candidate-gene approaches. Third, immigrant studies show that different migrating populations seem to respond in a similar way, at least in qualitative terms, implying that whatever differences exist in the genetic make up of the populations, the response to the western environment is largely similar, which is also consistent with an overall increase in mutagenic or mitogenic pressure. For example, Finns and Swedes have different population histories and gene pools[51,52]. The incidence of testicular cancer in Sweden is over double the rate in Finland. When Finns in their twenties move to Sweden, their testicular cancer risk remains at the Finnish level[53]. However, the risk in their sons equals the Swedish level, even if the mothers are Finnish and the sons' genotypes are completely Finnish.

The differences in cancer incidence between populations, and the large changes in incidence that occur over a relatively short period or on migration, are well established, and these constitute the most pertinent evidence for environmental causation of cancer. The criticisms that have been raised against the reliability of epidemiological research have not shaken the basis of this evidence[6,54]. However, these profound changes in incidence have attracted curiously little scientific attention, so the exact nature of the environmental causation and its possible modulation by genes remain largely unknown.

Common disease–common or rare variants

The shift of focus from Mendelian to complex diseases has prompted a new gene-identification strategy, according to which the classical experience with rare Mendelian diseases is contrasted with the 'common diseases–common variant hypothesis'[10,11,18,55–57] (BOX 2). The common disease–common variant hypothesis lends itself to genome-wide association studies that use the linkage between the disease allele and the marker allele (linkage disequilibrium) as a mapping tool; the Mendelian paradigm continues to emphasize family-based approaches, thereby focusing on a limited number of disease alleles[11,58].

Many complex diseases are characterized by a small Mendelian component, a somewhat larger familial (non-Mendelian) component and a large sporadic component. The genetic bases of many of the Mendelian components have been resolved using pedigree-based linkage studies. Among the most prevalent hereditary cancers, hereditary non-polyposis colorectal cancer (HNPCC) accounts for 1–3% of colorectal cancers[59,60], and *BRCA1* and *BRCA2* combined account for 2% of breast cancers[61]; for ovarian cancer the combined attributable fraction of *BRCA1* and *BRCA2* could be over 10% (REF. 62). However, the attributable fractions of hereditary syndromes depend on the frequency of the disease variants in the population, which can be highly variable. The figures that are given above refer to certain Western European and North American Caucasian populations.

A study of 1,150 cases of bladder cancer and a similar number of controls indicated that the genes N-acetyltransferase 2 (*NAT2*) and glutathione S-transferase M1 (*GSTM1*) explain 31% of cases[63]. The estimated odds ratios were 1.4 for the *NAT2* slow acetylator genotype (in reference to aromatic amines) and 1.7 for the *GSTM1* null genotype. The large population attributable fraction was due to the fact that over half of the population carried the risk alleles. Most cases were attributable to *GSTM1* in this male-dominated population; the relative risk of the *GSTM1* polymorphism was equally large in smokers and non-smokers, which was interpreted as equal protection against tobacco-related and non-tobacco-related carcinogenesis by the functional *GSTM1* gene. But, smoking alone is assumed to account for more than 60% of the population attributable fraction of male bladder cancer[30]. The significance of minimally increased odds ratios, such as

1.4, will probably continue to be debated until direct mechanistic evidence can be invoked. In the above study, the *NAT2* effect showed an interaction with smoking, and was more intense in smokers, giving credibility to the findings and supporting the predicted role of aromatic amines in bladder carcinogenesis[63].

Many initially positive associations of low-penetrance genes with disease have not been replicated when larger populations are analysed[17,64]. Even some variants that are presented as 'proof-of-principle' have failed in subsequent tests[9]. Five genes, *NAT2*, Harvey rat sarcoma virus oncogene 1 (*HRAS1*), glutathione S-transferase-θ 1 (*GSTT1*), tumour necrosis factor-α *(TNFA*) and 5,10-methylenetetrahydrofolate reductase (NADPH; *MTHFR*), have been reported to explain 54–64% of colorectal cancer, depending on the model that was applied[65]. Although the formal calculations for these results are correct, the moot question is whether these genes are related to colorectal cancer at all, and to what extent they can be replicated in large, ongoing studies[66]. The studies on these genes were conducted on patient populations that were not selected for family history, a design in which a large sample size is thought to compensate for the diluted familial effect. Genomic scientists should not forget the value of sampling in affected families[1] in which disease genes would be enriched. This approach would also force them to consider the extent of familial clustering as a likely measure of success.

A cancer clinician who is accustomed to seeing HNPCC patients would be surprised by the message that, in addition to the 1–3% of colorectal cancer patients with HNPCC, around 60% of his patients are suffering from a heritable disease that is caused by one of five genes (*HRAS1*, *NAT2*, *GSTT1*, *TNFA* and *MTHFR*) that he or she has never heard of. In fact, a geneticist who is working on colorectal cancer would be even more perplexed. If over 60% of the genetic causation were already known, he or she would have 'only' 40% to work towards. Little would be left to an environmental epidemiologist, who would hope to identify gene–environment interactions. We have already explained the reasons for such paradoxical claims, which are bound to become common in the common disease–common variant era. Strong evidence is needed to convince the scientific community of the validity of population attributable fractions of 30–60%, because the most prevalent known cancer genes, *BRCA1*, *BRCA2* (breast cancer) and mismatch repair genes (HNPCC) account for only about 1%. There are two teleological arguments against the idea of the five genes discussed above accounting for 60% of colorectal cancer. First, because colorectal cancer in the Western and Japanese populations is governed by environmental influences, it is unlikely that scientists will (ever) reach a consensus on which genes might account for 60% of the ill-defined heritable causation, even in a single population. Second, these five genes were identified about 20 years ago, with little direct mechanistic link to colorectal cancer. With the repertoire of 30,000 genes that are currently thought to exist in the human genome, such *a priori* success seems unlikely.

There is an important role for the population attributable fraction and familial risk in the *a posteriori* assessment of the likelihood of genetic effects, which we alluded to earlier. The familial risk of colorectal cancer is 1.8, and less than half of it is explained by known susceptibility genes[67,68]. The unexplained familial risk is therefore in the order of 1.5. Even if they are truly causative, the five genes that were discussed above would explain a familial risk of no more than about 1.1 (in an additive model), according to the calculation presented in TABLE 1. There is clearly a discrepancy: genes with a population attributable fraction of 60% account for only 20% of the familial risk. If the genes were ever to explain 100% of the disease in the population, all of the familial risk would need to be accounted for. The inconsistency between the population attributable fraction and the familial risk might be a third argument against the dominance of these five genes in colorectal carcinogenesis. The comparison of population attributable fractions and familial risks will be a test for the common disease–common variant hypothesis, because many of the suggested findings are likely to resemble the example of these five genes, in that they explain too much of the population attributable fraction but too little of the familial risk. The early literature on common candidate genes has examples

Box 2 | **Allelic architecture of complex diseases: contrasted models**

The haplotypes (sets of alleles on a single chromosome) of living individuals are inherited from ancestors, and they have been modified over generations through recombination events. The frequencies of the haplotypes have been governed by mutation rates, genetic drift, selection and population bottlenecks. The number of recombination events is related to the number of meioses. Therefore, close relatives share long-range haplotypes, over many haplotype blocks. These are DNA sequences with low rates of recombination. The stucture of haplotype blocks varies along the chromosomes and between populations; on average, haplotype blocks are of the order of 10 kb long and at each locus there are about 5 different blocks of variable frequency[10]. These data refer to the HapMap results, which were generated with 30 parent–offspring trios, enabling the detection of only the most common haplotypes. Family-based linkage studies can be carried out with a few hundred microsatellite markers because of the extensive haplotype sharing between family members. In SNP-based whole-genome association studies on outbred populations, individual haplotype blocks must be identified. To do this, several hundred thousand SNPs are required[10]. Among the crucial questions that remain to be answered regarding the architecture of disease alleles are the timing and frequency of mutations in the ancestral history. These have a bearing on the detection strategy that is used for disease genes, with opposing views:

Mendelian diseases

Variants that cause rare diseases have arisen independently on different ancestral haplotypes. Linkage studies in pedigrees might be preferable to association mapping, because a single disease haplotype would be expected in a family. All classical heritable traits conform to Mendelian inheritance and many common diseases have one or more Mendelian components. Genes that have been identified for many known Mendelian cancer syndromes are also mutated in sporadic cancers, including P53 (many sporadic cancers), identified in Li-Fraumeni syndrome, APC (colon cancer), identified in familial adenomatous polyposis and VHL (renal cancer), identified in von Hippel-Lindau disease[60].

Common disease–common variant hypothesis

Few common disease alleles fall within common haplotypes that are amenable to association mapping. These alleles could act jointly with other common susceptibility variants. Because of the low risk of each disease allele, family-based sampling offers only a limited advantage. The apoliprotein E allele ε4, which predisposes to Alzheimer and cardiovascular diseases, is a prime example of a common disease allele. A recent analysis of 871 candidate genes in lung cancer implicated many genes in the growth hormone–insulin like growth factor pathway[77]. The implication of many genes in a single pathway adds to the credibility of the findings.

of such excessive genotype odds ratios for single variants that they alone would account for more than 100% of the empirical familial risk[23]. High genotype odds ratios are rarely seen in the current literature; the ones that do appear probably raise as much suspicion as enthusiasm.

Small genetic risk

A now well-recognized problem of many early candidate-gene studies is the small sample size[69]. A recent trend has been to use ever larger sample sizes, allowing odds ratios of 1.4 and below to be called significant. The sample sizes that are required become many times larger when significance levels are adjusted for genome-wide comparisons, accommodating the concept that the tested genes are drawn from a pool of 30,000 genes, even if tested individually[11]. The value of an association study that uses 10,000 cases and 10,000 controls to find a gene that poses a risk of 1.3 might be questioned[11]. In sample-size calculations within the UK Biobank, case populations of 10,000 are also considered to detect risks of ≥ 1.15 for a single gene, and of 1.5–2.0 for interactions between two factors: genetic or environmental[9]. Interactions are one of the tenets of the multifactorial disease concept and it is important that they can be addressed in large studies.

The practical significance of a genetic risk below 1.5 is not obvious[18], although any reliable genetic data are of aetiological interest. Clinical counselling guidelines have been developed for high-penetrance cancer genes. Recommendations are available, even for prostate cancer, although no susceptibility genes have been identified[70]. Although the American Cancer Society Guidelines recommend certain actions for colorectal cancer when the familial risk is about 2.2 (REF. 71), no clinical genetic recommendations are available for many cancers that are rarer but have a higher familial risk.

Individualized medicine has been marked as one of the benefits of the Human Genome Project[72,73]. Accordingly, genomic tools are to be used to predict an individual's health and disease profiles, and his or her response to therapeutics. However, individual risks must be reasonably high before medical advice can be offered, in order to meet the principles of medical ethics[74]. It will be difficult to convey to individuals the practical benefits of informing them that he or she is a carrier of a common gene variant that confers a risk of 1.5 to disease X, when nothing can be said about other diseases that relate to this gene, nor about the ways of reducing the risk of disease X.

Conclusions

Whole-genome association studies are either ongoing or planned for many important diseases, with the belief that "…most diseases are the result of the interactions of multiple genes and environmental factors,"[1] and that the common disease–common variant hypothesis will turn out to be useful. However, the scientific arguments that are presented here for cancer aetiology, ranging from twin and family studies to migrant studies, as well as the vast incidence changes that occur over the course of one or two generations, demonstrate the unquestionable role of the complex environment. Failures and disappointments, even in the most advanced studies, might simply be due to the low heritability of the disease under study. Moreover, in many ongoing genomic studies, the environmental component is completely missing, indicating that genes, rather than gene–environment interactions, are assumed to be the cause[4,14]. In a recently announced funding scheme, the US National Institutes of Health are planning to implement a Genes and Environment Initiative, combining the analysis of genetic variation in patients with the development of environmental exposure monitoring.

Although we find little evidence that "…gene–environmental interactions underlie almost all cancers,"[2] the massive, ongoing efforts will undoubtedly detect some moderate-risk genes and truly link them to heritable diseases. For example, it would be surprising if no moderate-risk genes for breast and colorectal cancers were found, because the currently known high-risk genes only explain a proportion of the known familial risks. By the same token, we must remember that linkage studies have largely been negative for prostate cancer, even though large multinational resources of prostate cancer families have been used[75,76], although twin studies have suggested that prostate cancer has the highest heritability among common cancers[36]. The recently established susceptibility locus on chromosome 8, which has been confirmed in many populations, was initially implicated in an Icelandic linkage study, and might be evidence for the increased power of family-based studies that are carried out in homogeneous populations[76]. Genetic heterogeneity, many genes causing the same disease, is an inevitable problem in any disease identification strategy, which can be avoided when homogeneous populations are used.

There seems to be a surprising imbalance in scientific priorities: although, in some countries, efforts are mounted to resolve risks of 1.2, little attention is paid to the causes of the dramatic 10-fold increase in colon cancer that has occurred in Japan over the course of the past 30 years. Also neglected are the important incidence changes that have occurred elsewhere, among immigrants in the developed countries, and populations in the developing countries that are adopting western lifestyles. Understanding these macrochanges would teach us about the essence of complex diseases, and about the elusive gene–environment interactions, which might not be captured by the traditional environmental risk factors such as tobacco smoking. Genomics of diseases that are common in developed countries could probably be more effectively addressed in populations in which these diseases are rare, because the environmentally caused background contribution would be low in such populations. There is evidence from immigrant studies that these environmental factors are causing disease in adults, however, early childhood could be the period in which patterns for an individuals' risk of cancer are set[43,45,46]. If early life were the crucial period for gene–environment interactions, biobanks of adult blood samples and exposure information might have difficulty detecting them.

The call for ever-larger sample sizes seems to signal that the genetic effects in complex diseases are weaker than previously thought. Genetic risks below 1.5, although relevant for aetiological understanding, are without practical value, and the much-touted future of individualized genomic medicine seems to fade away as calculated genetic risks are seen to drop. An obvious benefit of large sample sizes will be the possibility to analyse some of the stronger binary interactions (gene–gene and gene–environment), in line with a complex disease paradigm. The interacting variants are likely to occur infrequently in the population, bringing complex disease back to the realm of rare variants, in which genomic medicine could find a niche.

Kari Hemminki and Asta Försti are at the Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany, and the Center for Family Medicine, Karolinska Institute, 141 83 Huddinge, Sweden.

Justo Lorenzo Bermejo is at the Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ).

Correspondence to K.H.
e-mail: k.hemminki@dkfz.de

doi:10.1038/nrg2009

1. Guttmacher, A. E., Collins, F. S. & Carmona, R. H. The family history — more important than ever. *N. Engl. J. Med.* **351**, 2333–2336 (2004).
2. Khoury, M. J., Davis, R., Gwinn, M., Lindegren, M. L. & Yoon, P. Do we need genomic research for the prevention of common diseases with environmental causes? *Am. J. Epidemiol.* **161**, 799–805 (2005).
3. Doll, R. & Peto, R. The causes of cancer. *J. Natl Cancer Inst.* **66**, 1191–1308 (1981).
4. Willett, W. Balancing life-style and genomics research for disease prevention. *Science* **296**, 695–698 (2002).
5. Merikangas, K. R. & Risch, N. Genomic priorities and public health. *Science* **302**, 599–601 (2003).
6. Buchanan, A. V., Weiss, K. M. & Fullerton, S. M. Dissecting complex disease: the quest for the Philosopher's Stone? *Int. J. Epidemiol.* **35**, 562–571 (2006).
7. Baker, S. G. & Kaprio, J. Common susceptibility genes for cancer: search for the end of the rainbow. *BMJ* **332**, 1150–1152 (2006).
8. Hakama, M. Family history in colorectal cancer surveillance strategies. *Br. Med. J.* **368**, 101–103 (2006).
9. Davey Smith, G. *et al.* Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* **366**, 1484–1498 (2005).
10. The International Hap Map Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
11. Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
12. Risch, N. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
13. Peltonen, L. & McKusick, V. Dissecting human diseases in the postgenomic era. *Science* **291**, 1224–1249 (2001).
14. Vineis, P. A self-fulfilling prophecy: are we underestimating the role of environment in gene-environment interaction research? *Int. J. Epidemiol.* **33**, 945–946 (2004).
15. Colditz, G. A., Sellers, T. A. & Trapido, E. Epidemiology — identifying the causes and preventability of cancer? *Nature Rev. Cancer* **6**, 75–83 (2006).
16. Caspi, A. & Moffitt, T. E. Gene–environment interactions in psychiatry: joining forces with neuroscience. *Nature Rev. Neurosci.* **7**, 583–590 (2006).
17. Burton, P., Tobin, M. & Hopper, J. Key concepts in genetic epidemiology. *Lancet* **366**, 941–951 (2005).
18. Terwilliger, J. D. & Weiss, K. M. Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann. Med.* **35**, 532–544 (2003).
19. Hopper, J. L., Bishop, D. T. & Easton, D. F. Population-based family studies in genetic epidemiology. *Lancet* **366**, 1397–1406 (2005).
20. Peto, J. & Houlston, R. Genetics and the common cancers. *Eur. J. Cancer* **37**, S88–S96 (2001).
21. Risch, N. The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol. Biomarkers Prev.* **10**, 733–741 (2001).
22. Wang, W. Y. S., Cordell, H. J. & Todd, J. A. Association mapping of complex diseases in linked regions: estimation of genetic effects and feasibility of testing rare variants. *Genet. Epidemiol.* **24**, 36–42 (2003).
23. Hemminki, K., Försti, A. & Lorenzo Bermejo, J. Single nucleotide polymorphisms (SNPs) are inherited from parents and they measure heritable events. *J. Carcinog.* **4**, 2 (2005).
24. Hemminki, K. & Lorenzo Bermejo, J. Relationships between familial risks of cancer and the effects of heritable genes and their SNP variants. *Mutat. Res.* **592**, 6–17 (2005).
25. Hemminki, K. & Lorenzo Bermejo, J. Constraints for genetic association studies imposed by attributable fraction and familial risk. *Carcinogenesis* 28 Sep 2006 [epub ahead of print].
26. Hemminki, K. & Li, X. Familial risks of cancer as a guide to gene identification and mode of inheritance. *Int. J. Cancer* **110**, 291–294 (2004).
27. Chen, S. *et al.* Characterization of *BRCA1* and *BRCA2* mutations in a large United States sample. *J. Clin. Oncol.* **24**, 863–871 (2006).
28. Vogel, F. & Motulsky, A. *Human Genetics: Problems and Approaches* (Springer, Heidelberg, 1996).
29. Higginson, J. Present trends in cancer epidemiology. *Proc. Can. Cancer Conf.* **8**, 40–75 (1968).
30. IARC. *Tobacco Smoke and Involuntary Smoking* (IARC, Lyon, 2004).
31. IARC. *Cancer: Causes, Occurence and Control* (IARC, Lyon, 1990).
32. Vogelstein, B. & Kinzler, K. *The Genetic Basis of Human Cancer* (McGraw–Hill, New York, 2002).
33. Rothman, K. & Greenland, S. *Modern Epidemiology* (Lippincott–Raven, Philadelphia, 1998).
34. Ponder, B. Cancer genetics. *Nature* **411**, 336–341 (2001).
35. Hemminki, K. & Czene, K. Attributable risks of familial cancer from the Family-Cancer Database. *Cancer Epidemiol. Biomarkers Prev.* **11**, 1638–1644 (2002).
36. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer. *N. Engl. J. Med.* **343**, 78–85 (2000).
37. Stewart, B. & Kleihues, P. (eds) *World Cancer Report* 351 (IARC, Lyon, 2003).
38. Amundadottir, L. T. *et al.* Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family. *PLoS Med.* **1**, e65 (2004).
39. Kerber, R. A. & O'Brien, E. A cohort study of cancer risk in relation to family histories of cancer in the Utah population database. *Cancer* **103**, 1906–1915 (2005).
40. Soffritti, M. *et al.* First experimental demonstration of the multipotential carcinogenic effects of aspartame administered in the feed to Sprague–Dawley rats. *Environ. Health Perspect.* **114**, 379–385 (2006).
41. Doll, R., Peto, R., Boreham, J. & Sutherland, I. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* **328**, 1519 (2004).
42. IARC. *Cancer Incidence in Five Continents* (IARC, Lyon, 2002).
43. Parkin, D. M. & Khlat, M. Studies of cancer in migrants: rationale and methodology. *Eur. J. Cancer* **32A**, 761–771. (1996).
44. McCredie, M. Cancer epidemiology in migrant studies. *Recent Results Cancer Res.* **154**, 298–305 (1998).
45. Hemminki, K., Li, X. & Czene, K. Cancer risks in first-generation immigrants to Sweden. *Int. J. Cancer* **99**, 218–228 (2002).
46. Hemminki, K. & Li, X. Cancer risks in second-generation immigrants to Sweden. *Int. J. Cancer* **99**, 229–237 (2002).
47. The National Board of Health and Welfare Center for Epidemiology. *Cancer Incidence in Sweden 2002* [online], < www.sos.se > (2002).
48. Polednak, A. Trends in cancer incidence in Connecticut, 1935–1991. *Cancer* **74**, 2863–2872 (1994).
49. Yiu, H. Y., Whittemore, A. S. & Shibata, A. Increasing colorectal cancer incidence rates in Japan. *Int. J. Cancer* **109**, 777–781 (2004).
50. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4**, 177–183 (2004).
51. Kittles, R. A. *et al.* Dual origins of Finns revealed by Y chromosome haplotype variation. *Am. J. Hum. Genet.* **62**, 1171–1179 (1998).
52. Sajantila, A. & Pääbo, S. Language replacement in Scandinavia. *Nature Genet.* **11**, 359–360 (1995).
53. Hemminki, K. & Li, X. Finnish and Swedish genotypes and risk of cancer in Sweden. *Eur. J. Hum. Genet.* **11**, 207–209 (2003).
54. von Elm, E. & Egger, M. The scandal of poor epidemiological research. *Br. Med. J.* **329**, 868–869 (2004).
55. Weiss, K. M. & Terwilliger, J. D. How many diseases does it take to map a gene with SNPs? *Nature Genet.* **26**, 151–157 (2000).
56. Terwilliger, J. On the resolution and feasibility of genome scanning approaches. *Adv. Genet.* **42**, 351–391 (2001).
57. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease–common variant … or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
58. Terwilliger, J. D. & Hiekkalinna, T. An utter refutation of the 'Fundamental Theorem of the HapMap'. *Eur. J. Hum. Genet.* **14**, 426–437 (2006).
59. Lynch, H. T. & de la Chapelle, A. Hereditary colorectal cancer. *N. Engl. J. Med.* **348**, 919–932 (2003).
60. Nagy, R., Sweet, K. & Eng, C. Highly penetrant hereditary cancer syndromes. *Oncogene* **23**, 6445–6470 (2004).
61. Syrjäkoski, K. *et al.* Population-based study of *BRCA1* and *BRCA2* mutations in 1,035 unselected Finnish breast cancer patients. *J. Natl Cancer Inst.* **92**, 1529–1531 (2000).
62. Risch, H. A. *et al.* Prevalence and penetrance of germline *BRCA1* and *BRCA2* mutations in a population series of 649 women with ovarian cancer. *Am. J. Hum. Genet.* **68**, 700–710 (2001).
63. Garcia-Closas, M. *et al. NAT2* slow acetylation, *GSTM1* null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet* **366**, 649–659 (2005).
64. Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nature Genet.* **29**, 306–309. (2001).
65. Yang, Q., Khoury, M., Friedman, J., Little, J. & Flanders, W. How many genes underlie the occurence of common complex diseases in the population? *Int. J. Epidemiol.* **34**, 1129–1137 (2005).
66. Webb, E. L. *et al.* Search for low penetrance alleles for colorectal cancer through a scan of 1467 non-synonymous SNPs in 2,575 cases and 2,707 controls with validation by kin-cohort analysis of 14,704 first-degree relatives. *Hum. Mol. Genet.* **15**, 3263–3271 (2006).
67. Hemminki, K. & Chen, B. Familial risk for colorectal cancers are mainly due to heritable causes. *Cancer Epidemiol. Biomarkers Prev.* **13**, 1253–1256 (2004).
68. Johns, L. E. & Houlston, R. S. A systematic review and meta-analysis of familial colorectal cancer risk. *Am. J. Gastroenterol.* **96**, 2992–3003 (2001).
69. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl Cancer Inst.* **96**, 434–442 (2004).
70. ASCO. American Society for Clinical Oncology policy statement update: genetic testing for cancer susceptibility. *J. Clin. Oncol.* **21**, 2397–2406 (2003).
71. Hemminki, K., Li, X. & Czene, K. Familial risk of cancer: data for clinical counseling and cancer genetics. *Int. J. Cancer* **108**, 109–114 (2004).
72. Collins, F. S. & McKusick, V. Implications of the Human Genome Project for medical science. *JAMA* **285**, 540–544 (2001).
73. Guttmacher, A. E. & Collins, F. S. Welcome to the genomic era. *N. Engl. J. Med.* **349**, 996–998 (2003).
74. Burke, W. & Press, N. Genetics as a tool to improve cancer outcomes: ethics and policy. *Nature Rev. Cancer* **6**, 476–482 (2006).
75. Xu, J. *et al.* A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. *Am. J. Hum. Genet.* **77**, 219–229 (2005).
76. Amundadottir, L. T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nature Genet.* **38**, 652–658 (2006).
77. Rudd, M. F. *et al.* Variants in the GH–IGF axis confer susceptibility to lung cancer. *Genome Res.* **16**, 693–701 (2006).

Acknowledgments

Supported by Deutsche Krebshilfe, Swedish Cancer Society, Swedish Council for Working Life and Social Research and the EU.

Competing interests statement

The authors declare no competing financial interests.

DATABASES

The following terms in this article are linked online to:

Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

BRCA1 | *BRCA2* | *GSTM1* | *GSTT1* | *HRAS1* | *MTHFR* | *NAT2* | *TNFA*

UniProtKB: http://ca.expasy.org/sprot

P53

FURTHER INFORMATION

American Cancer Society Guidelines: http://www.cancer.org/docroot/PED/content/PED_2_3X_ACS_Cancer_Detection_Guidelines_36.asp

Connecticut Tumor Registry: http://www.dph.state.ct.us/oppe/hptumor.htm

German Cancer Research Center Division of Molecular Genetic Epidemiology homepage: http://www.dkfz.de/en/molgen_epidemiology/index.html

Human Genome Project: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

International Agency for Research on Cancer: http://www.iarc.fr

National Institutes of Health: http://www.nih.gov

NIH Genes and Environment Initiative: http://www.gei.nih.gov

Swedish Cancer Registry: http://www.sos.se/epc/english/cancereng.htm

UK Biobank: http://www.ukbiobank.ac.uk

Access to this links box is available online.