Review

The Genetic Epidemiology of Cancer: Interpreting Family and Twin Studies and Their Implications for Molecular Genetic Approaches

Neil Risch[1]

Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120, and Division of Research, Kaiser Permanente, Oakland, California 94611-5714

Abstract

The recent completion of a rough draft of the human genome sequence has ushered in a new era of molecular genetics research into the inherited basis of a number of complex diseases such as cancer. At the same time, recent twin studies have suggested a limited role of genetic susceptibility to many neoplasms. A reappraisal of family and twin studies for many cancer sites suggests the following general conclusions: (*a*) all cancers are familial to approximately the same degree, with only a few exceptions (both high and low); (*b*) early age of diagnosis is generally associated with increased familiality; (*c*) familiality does not decrease with decreasing prevalence of the tumor–in fact, the trend is toward increasing familiality with decreasing prevalence; (*d*) a multifactorial (polygenic) threshold model fits the twin data for most cancers less well than single gene or genetic heterogeneity-type models; (*e*) recessive inheritance is less likely generally than dominant or additive models; (*f*) heritability decreases for rarer tumors only in the context of the polygenic model but not in the context of single-locus or heterogeneity models; (*g*) although the family and twin data do not account for gene-environment interactions or confounding, they are still consistent with genes contributing high attributable risks for most cancer sites. These results support continued search for genetic and environmental factors in cancer susceptibility for all tumor types. Suggestions are given for optimal study designs depending on the underlying architecture of genetic predisposition.

Introduction

Last June, human genome scientists announced completion of and more recently published a rough draft of the human genome sequence, ushering in a new era of human molecular genetics (1, 2). This accomplishment was heralded with great fanfare and with predictions of a significant impact on the understanding and treatment of chronic human diseases such as cancer. It is perhaps for this reason that a recent twin study of cancer (3) received so much attention from both the scientific and lay media, because its conclusion was that susceptibility to cancer is primarily environmental and not inherited and thus seemingly at odds with the claims of the genomicists. The fact that this was by far the largest twin study in cancer yet reported (nearly 45,000 twin pairs) also lent credence to this conclusion, although an editorial appearing in the same issue (4) discussed some of the limitations of that study. Given the increasing emphasis on molecular genetic approaches to address familial disorders coupled with the latest evidence questioning the role of genetics in cancer susceptibility, a reassessment of the role of genetic factors in cancer susceptibility generally and for site-specific cancers in particular appears warranted.

Study Designs

Familial aggregation of a trait is a necessary but not sufficient condition to infer the importance of genetic susceptibility, because environmental and cultural influences can also aggregate in families, leading to family clustering and excess familial risk. Family aggregation is usually assessed by studying relatives of affected subjects and contrasting their rates of illness with those of a suitable control group, typically the relatives of unaffected subjects.

Several approaches for disentangling genetic from environmental influences are also possible in studies of human disease, although practical difficulties often limit their use. The most powerful design examines risks in biological relatives of affected *versus* control adoptees, because adoption creates a separation between an individual's biological and environmental influences. Because it is often difficult to obtain access to information on biological relatives of adoptees, adoption studies typically focus only on common disease or trait outcomes.

Another study design often used to separate genetic and environmental influences involves twins. Identical (MZ[2]) twins derive from the fission of a single fertilized egg and thus inherit identical genetic material. By contrast, fraternal (DZ) twins derive from two distinct fertilized eggs and thus have the same genetic relationship as full siblings, although they may be more "biologically" related because of sharing the same prenatal intrauterine experience.

Comparing the similarity of MZ twins with same-sex DZ twins is a common approach for gleaning the magnitude of genetic influence on a disease or trait and has been applied extensively to a broad range of disorders, including cancer. A standard measure of similarity used in twin studies is the concordance rate. The "pairwise" concordance is calculated simply as the proportion of twin pairs with both twins affected of all ascertained twin pairs with at least one affected. On the other hand, the "probandwise" concordance allows for double counting of doubly ascertained twin pairs and has the advantage of being interpretable as the recurrence risk in a co-twin of an affected individual (5). Usually, the most critical assumption in

Received 2/1/00; revised 4/27/01; accepted 5/2/01.

[1] To whom requests for reprints should be addressed, at Department of Genetics, M322, Stanford University School of Medicine, Stanford, CA 94305-5120. Fax: (650) 725-1534; E-mail: risch@lahmed.stanford.edu.

[2] The abbreviations used are: MZ, monozygotic; DZ, dizygotic; MFT, multifactorial threshold; FRR, family risk ratio; SIR, standardized incidence ratio; PAF, population attributable fraction; RR, relative risk.

twin studies is that MZ and DZ twins display a comparable degree of similarity because of the sharing of environmental factors, so that the difference in concordance rates between MZ and DZ twins is only a reflection of genetic factors.

Table 1 MZ and DZ twin relative risks as a function of prevalence (K) and heritability (H) according to a multifactorial threshold model

K	H = 20%			H = 30%			H = 40%		
	λ_M	λ_D	R_{MD}	λ_M	λ_D	R_{MD}	λ_M	λ_D	R_{MD}
3%	2.4	1.6	2.4	3.5	2.0	2.5	4.9	2.4	2.7
1%	3.4	1.9	2.6	5.5	2.6	2.9	8.6	3.4	3.2
0.1%	6.8	2.8	3.3	15.0	4.6	3.9	29.6	6.8	4.9

Genetic Models and the Interpretation of Family and Twin Studies

Understanding empirical evidence about genetic susceptibility to cancer requires a discussion of models of genetic inheritance and their implications. The simplest way to measure genetic effects is through familial risk ratios, defined as the risk to a given type of relative of an affected individual divided by the population prevalence. These risk ratios have been given the notation λ (6), and specifically λ_S for the sibling risk ratio, λ_O for the offspring risk ratio, λ_P for the parent risk ratio, λ_1 for all first degree relatives combined (parents + siblings + offspring), λ_D for DZ twins, and λ_M for MZ twins.

If genetic susceptibility is attributable to a single (rare) dominant gene, it is easy to show (6) that $\lambda_P = \lambda_O = \lambda_S = \lambda_D = (\lambda_M + 1)/2$, which implies that the MZ:DZ ratio defined by $R_{MD} = (\lambda_M - 1)/(\lambda_D - 1) = 2$. On the other hand, if susceptibility is attributable to a recessive gene, $\lambda_P = \lambda_O < \lambda_S = \lambda_D$, with the degree of difference between λ_S and λ_O depending on the frequency of the "at-risk" allele (λ_S/λ_O ranging from near 1 for a very common allele to infinity for a very rare allele). For a recessive model, R_{MD} is usually >2, again depending on the allele frequency. For a rare allele, $R_{MD} = 4$, but diminishes toward 2 if the allele is very common.

How are these expectations altered if there are also nongenetic cases mixed in, or if more than one locus contributes to susceptibility? Nongenetic cases (or "phenocopies") do not influence the predictions given above. On the other hand, if more than one gene exists that influences susceptibility, the predictions may be altered, depending on whether interaction effects exist among the contributing genes ("epistasis" in genetics parlance). Specifically, if mutant alleles at different loci are individually rare, so that it is very unlikely that an individual would carry more than one (a scenario typically termed "genetic heterogeneity" or "locus heterogeneity" by geneticists), the same predictions as given above hold. Equivalently, for more common alleles, if the risk associated with carrying multiple mutants is additive, the same predictions hold (6). By contrast, if the risk associated with carrying multiple "at-risk" alleles is not additive, *e.g.*, multiplicative, a different pattern for the λ values than described above occurs. Specifically, R_{MD} is now >2 and can achieve very high values, depending on how many loci are involved and the degree of interaction.

Another genetic model that is commonly used in the analysis of family and twin data is the polygenic or MFT model. This model postulates a genetic basis consisting of numerous small, additive effects underlying a continuously distributed trait termed liability. The assumptions of the model allow invocation of a Gaussian distribution because of the Central Limit Theorem. It is further assumed that risk, as a function of liability, increases sigmoidally (asymptotically to 0 for liability equal to minus infinity and to 1 for liability equal to plus infinity). This sigmoid risk function is assumed to take the form of a cumulative normal distribution function. It can be shown that the latter assumption is mathematically equivalent to assuming an independent, additive random environmental component to liability, with a threshold imposed on the total liability scale determining affected status (*e.g.*, a total liability value above a threshold T implies affected, and below T unaffected). Thus, according to the MFT model, there are two additive, normally distributed components to liability, a polygenic component and a random environmental component. The proportion of the total variance of liability attributable to the polygenic component is usually termed "heritability," where it is understood that this refers to the heritability of the latent liability trait. Because it is based on the underlying liability variable, heritability is independent of the threshold T.

What are the implications of the MFT model for familial relative risks? The MFT model has two parameters, the heritability H (defined above) and threshold T, determined by the population prevalence K. For example, a trait with a prevalence K of 1% corresponds to a threshold T of 2.33 (in SD units above the mean 0). Two relatives are assumed to have a bivariate normal joint distribution of liability with correlation $\rho = rH$, where r is the coefficient of relationship for the relatives (r = 1 for MZ twins, r = 0.5 for first-degree relatives, r = 0.25 for second-degree relatives, and so on). The recurrence risk is then calculated as the joint probability K_2 that the liability values for both relatives exceed T, divided by K (the probability that the index relative's liability exceeds T). Then the familial risk ratio λ is given by K_2/K^2.

There are two important implications of the MFT model for the λ values. The first implication is that, for a fixed value of λ, the corresponding heritability H decreases with decreasing K. For example, for two traits, each with a λ_S value of 2, one with a prevalence of 10% will have a higher heritability estimate than one with a prevalence of 1%. The second implication of the MFT model is that the MZ:DZ ratio R_{MD} is always >2, indicating nonadditivity of gene effects. In fact, R_{MD} increases directly with heritability H and inversely with prevalence K. At first, this feature may seem contradictory to the basic assumption of the MFT model, *i.e.*, that the individual gene effects are additive. However, the additivity of gene effects is on the scale of liability. Because disease risk is not a linear function of liability but is sigmoidally related, on the risk scale the gene effects are nonadditive and thus give rise to interactive (or epistatic) effects. This is entirely analogous to the situation in epidemiology of defining whether interactions occur in the context of an additive or multiplicative model. In this case, in terms of recurrence risk patterns in relatives, it is nonadditivity on the *risk* scale (as opposed to liability scale) that matters.

These two characteristics of the MFT model are depicted in Table 1. As is evident there, both λ_M and λ_D increase dramatically with decreasing K for a fixed value of H, as does R_{MD}, especially for higher values of H. One can also see from this table that for a trait with $\lambda_D = 2$, the estimate of H ranges from 30% when K = 3% to 12% when K = 0.1%.

One may also ask how these predictions of λ_M and λ_D from genetic models are influenced by assumptions regarding environmental risk factors. The models above generally assume that environmental exposures are randomly distributed within families (or twins), thus inducing no additional familial or twin correlation. For exposures that do not cluster in families, the predictions given above hold true no matter what the relationship between genotype and exposure (*i.e.*, in the presence or absence of gene-environment interactions), because exposures are independent among family members. On the other hand, for exposures that are not randomly distributed in families, there will be an impact on familial recurrence, *i.e.*, to increase the λ values given above. The degree of increase depends on the frequency of exposure and extent of correlation for the exposure among family members.

With regard to twin studies, we need to consider the relative impact on MZ *versus* DZ twins. If the environmental exposure is correlated to a similar degree between MZ and DZ twins (a common assumption), the MZ:DZ ratio R_{MD} will actually be attenuated. This can be seen most simply by adding the same constant c to each of the twin risk ratios λ_{M} and λ_{D}. Then if $R_{\mathrm{MD}} = (\lambda_{\mathrm{M}} - 1)/(\lambda_{\mathrm{D}} - 1) = 2$ without the environmental correlation, with it $R_{\mathrm{MD}}' = (\lambda_{\mathrm{M}} - 1 + c)/(\lambda_{\mathrm{D}} - 1 + c) = 2 - c/(\lambda_{\mathrm{D}} - 1 + c) < 2$. On the other hand, if environmental exposure is correlated to a greater extent in MZ than DZ twin pairs, any result is possible, depending on the degree of difference. In particular, if the difference is large, R_{MD} may increase (>2), whereas if it is modest, R_{MD} may stay the same or decrease.

In general, the greatest opportunity for confounding arises for rare, powerful exposures that cluster strongly in families. Common or universal exposures are less likely to induce significant familial clustering.

Evidence of Familiality

Numerous studies have addressed the degree to which site-specific cancers run in families, *e.g.*, those of the breast, colon, prostate, and lung. However, few large-scale studies using a single standardized approach to many different cancer sites have been reported. Such studies are useful to derive a global view of the familiality of cancer. One such study from Utah examined the familial recurrence of cancer for 28 specific sites, based on 35,228 probands with cancer (7). These authors matched the Utah Genealogic Database, which provides names of all first-degree relatives of these probands (parents, siblings, and offspring), with the Utah Cancer Registry to determine the frequency of cancer in the first-degree relatives of these cases. Comparable expected rates for each cancer site were obtained from the 399,786 first-degree relatives of a matched control group. Familiality (FRR) was assessed as the ratio of the observed number of cancer cases among the first degree relatives of the probands divided by the expected number derived from the control relatives, based on the years of birth (cohort) of the case relatives. In essence, the FRR provides an age-adjusted risk ratio to first-degree relatives of cases compared with the general population and is thus equivalent to the parameter λ_1 defined above. These authors also examined family recurrence for a separate group of early-onset cancer probands (diagnosis prior to age 50 for melanoma, breast and brain/central nervous system cancers and prior to age 60 for all other cancers).

A second large population based study of family recurrence in Sweden for a variety of cancer sites has been reported recently (8). These authors studied cancer recurrence in >2 million nuclear families that were linked to the Swedish Cancer Registry. Specifically, they calculated SIRs for: (*a*) offspring with a parent but no sibling with cancer; (*b*) offspring with a sibling but no parent with cancer; and (*c*) offspring with both a parent and sibling with cancer. Among 4,225,232 parents, 435,000 (10.3%) had a diagnosis of cancer. The offspring were born after 1934 and followed up to 1996 and thus were between ages 0 and 61 at time of study. Among 5,520,756 offspring, 71,424 (1.3%) had a diagnosis of cancer. Among male offspring, the average age at diagnosis was 38; for female offspring, the mean age of diagnosis was 42.

Table 2 FRR for first-degree relatives of cancer probands by site, in decreasing order of prevalence

Data from Utah [Goldgar *et al.* (7)] and Sweden [Dong and Hemminki (8)] are shown.

Site	Utah		Sweden	
	FRR (total)	FRR (early onset)	FRR (Child)	FRR (sibling)
Prostate	2.21	4.08	2.82	9.41
Breast	1.83	3.70	1.86	2.01
Colorectal	2.54	4.53	1.86	4.41
Lung	2.55	2.50	1.68	3.16
Uterine	1.32	1.75	—	—
Melanoma	2.10	6.43	2.50	3.41
Bladder	1.53	5.00	1.53	3.30
Non-Hodgkin's lymphoma	1.68	2.40	1.68	2.37
Brain/CNS	1.97	8.95	1.72	2.37
Cervix	1.73		1.93	2.39
Ovary	2.04		2.94	2.52
Stomach	2.08		1.72	8.82
Lip	2.72			
Pancreas	1.25			
Kidney	2.46		1.60	5.26
Oral cavity	1.82			
Thyroid	8.48		9.51	12.42
Multiple myeloma	4.29		4.25	5.62
Granulocytic leukemia	2.94		1.69	3.53
Lymphocytic leukemia	5.00			
Hodgkin's lymphoma	1.25			
Female genitalia	2.22		2.85	3.97
Soft tissue sarcoma	2.00			
Testicular	8.57		4.31	8.50
Gallbladder	2.22			
Larynx	8.00			
Total, mean	2.12	3.78	2.14	3.37
Total, median	2.15	4.08	1.86	3.53

Results of analyses of both data sets are reproduced in Table 2 (colon, rectum, and anus have been combined into one site, colorectal, giving 26 total sites). An important question is whether the rare cancer sites are less familial than the common sites. Thus, in Table 2 the various sites are listed in decreasing order of prevalence as reported in Utah. FRRs are given for all and early-onset probands as reported by Goldgar *et al.* (7) for Utah. For the Swedish data, all offspring with an affected parent (*i.e.*, groups in *a* and *c* defined above) were combined to obtain a complete estimate of the offspring recurrence risk (SIR) and similarly groups in *b* and *c* were combined to obtain an overall sibling recurrence risk (SIR). Hence, the numbers provided in Table 2 differ somewhat from those provided in the original tables from the Swedish study (8) but are more directly comparable with the figures from Utah.

There are three important observations to be derived from the numbers in Table 2:

(*a*) There is remarkable similarity of the FRRs across cancer sites, with a mean value (weighted by prevalence) of 2.12 and median of 2.15 for Utah (all probands) and a mean and median of 2.14 and 1.86, respectively, for the Swedish offspring. There are a few notable exceptions, however. Thyroid, testicular, and laryngeal cancers and lymphocytic leukemia and multiple myeloma appear to have elevated recurrence risks in both studies. Also in both studies, all FRRs are >1, and 18 of the 26 sites (Utah) and 14 of 17 sites (Sweden) have an FRR value between 1.5 and 3.0. Furthermore, there is overall consistency in the FRRs from Utah (all probands) and Sweden (offspring). For the 17 sites in common between the two studies, the correlation in FRRs is 0.83. However, this correlation is primarily attributable to the high values for thyroid, melanoma, and testis in both studies. After removing these 3 sites, the correlation becomes 0.01. This observation probably reflects a lack of true variation around the average FRR of 2 for the remaining sites, the observed variation being primarily random (*i.e.,* statistical noise).

(*b*) It is apparent from Table 2 that there is no decline in FRR with decreasing frequency of the cancer site. In fact, if anything, there is a trend toward increasing FRR with decreasing frequency. For example, for Utah, for the first 13 cancers listed in Table 2, the (weighted) average FRR is 2.08 (median, 2.04), whereas for the second 13 cancers, the (weighted) average FRR is 3.29 (median, 2.46). Similarly, in Sweden, for the first 9 sites listed, the average (median) FRR for offspring is 1.94 (1.86), whereas for the latter group of 9 sites, the average (median) is 3.39 (2.85). Thus, when characterized by FRR, rarer cancers are no less familial (and probably more familial) than the common cancers. They may appear "sporadic" because they are rare and most often occur in the absence of a family history (*i.e.,* families with multiple cases are rare). However, when assessed systematically, relatives of cases with rare cancers have at least the same degree of increased risk (or more) compared with the relatives of cases with common cancers.

(*c*) Table 2 reveals increased family recurrence associated with early age at diagnosis. For Utah, for the nine cancer sites listed, the (weighted) average FRR for the early onset probands was 3.78 (median, 4.08). This figure is nearly 2-fold greater than the FRR for the same 9 sites for all probands (average, 2.08; median, 1.97). Eight of the nine sites listed showed an increase in FRR with early onset (only lung did not). Thus, it appears to be a generalizable conclusion that increased familiality is associated with early age of diagnosis.

In Sweden, the authors did not separate out their data based on age of onset of the proband. Rather, they calculated familial recurrence for offspring with an affected parent and for offspring with an affected sibling. As can be seen in Table 2, the sibling recurrence ratios (mean, 3.37; median, 3.53) are systematically higher than the offspring recurrence ratios (mean, 2.14; median, 1.86). The authors interpreted the elevated sibling recurrence risk ratios as evidence of recessive gene action, because recessive genes lead to increased risk in siblings compared with offspring (8). However, this conclusion is completely confounded by the fact that for the offspring with affected siblings, the average age of diagnosis of the affected sibling was only 38–42 on average, much younger than the average age of diagnosis of the affected parent for offspring with an affected parent (probably by 30 years or so). This was attributable to the fact that offspring were young at the time of study (maximum age, 61), whereas the parents were not. This is also reflected in the vastly lower prevalence among the offspring (siblings), 1.3%, compared with the parents, 10.3%. Thus, it is more likely that the elevated risks in siblings *versus* offspring of cancer probands as observed in the Swedish data are a reflection of increased familial risk being associated with early age at diagnosis, as also seen in Utah, rather than with recessive genes.

Separating Genes from Environment–Adoptees and Twins

Because of the difficulty in conducting adoption studies, only one such study in cancer has been reported (9). This study examined cancer mortality in the 593 biological (but adopted away) offspring of parents who died of cancer (all sites combined) by age 50. A hazard ratio of 1.2 for cancer mortality in the child (all sites) was observed, although this figure was not significantly >1. The number of subjects was far too few to examine individual cancer sites. The implication of studying all cancer sites pooled is discussed further below.

With respect to twin studies, because most cancers are rare and occur late in life, large twin cohorts are usually required to obtain sufficient cases. Thus, few twin studies in cancer have been reported, and these have focused primarily only on the commonly occurring cancer sites or on all sites combined.

For example, the National Academy of Sciences Twin Cohort, containing nearly 16,000 male veteran twins, revealed no increased concordance in lung cancer mortality in MZ *versus* DZ twins (despite an observed increase in concordance for cigarette smoking in the MZ twins). This led the authors to conclude that genetic susceptibility has little influence on lung cancer mortality (10). The same cohort was also studied for death from all cancer sites combined (11). Here the ratio of MZ:DZ concordance was 1.4, modestly suggestive of genetic influence. This cohort was also evaluated for prostate cancer risk (12). In this case, the MZ concordance was estimated to be 27.1% compared with 7.1% for DZ twins, giving a concordance ratio of 3.8, strong evidence for the influence of genetic susceptibility. The authors estimated the heritability of liability (described above) to be 57%.

Instead of large, population-based twin cohorts, an alternative strategy is to identify twins from a large sample of cancer cases and follow their co-twins for their cancer risk. This clinic-based approach was used to study Hodgkin's lymphoma (13), where 366 (179 MZ and 187 DZ) twins with the disease were identified, and their co-twins were followed. Ten of the 179 MZ co-twins similarly developed Hodgkin's disease, compared with none of the 187 DZ co-twins, suggesting a strong heritable component to this form of cancer.

The Nordic countries are an ideal setting for population-based twin studies because of the existence of population-based twin and cancer registries. For example, two Swedish twin cohorts, one born between 1886 and 1925 with 10,503 pairs and another born between 1926 and 1958 with 12,883 pairs were linked to that nation's cancer registry (14). The authors found increased concordance in MZ *versus* DZ twins for colorectal, breast, cervical, and prostate cancers, suggesting the importance of genetic factors for these sites; by contrast, MZ and DZ concordance were comparable for stomach and lung cancers, suggesting less of a genetic role in these cancers.

Similarly, in Finland, 12,941 same-sex twin pairs were linked to that country's cancer registry (15). Examining all sites combined, these authors estimated a low overall influence of genetic factors (heritability of 18%) and thus concluded that the environment plays the major role in cancer susceptibility.

Most recently, the twin registry of Denmark was linked to that nation's cancer registry and combined with similar analyses in Sweden and Finland to produce the largest population-based twin study of cancer to date (3). In total, 44,788 same-sex twins were followed for cancer prevalence. Because of modest heritability estimates, the authors concluded that "inherited genetic factors make a minor contribution to susceptibility to most types of neoplasms." Because of the size of this study and

Table 3 Twin site-specific RRs for cancer

Sex and site	K	λ_M	λ_D	R_{MD}
Male				
Prostate	2.4%	8.06	2.83	3.86
Lung	1.7%	6.27	6.14	1.03
Colon	1.5%	5.87	5.14	1.18
Stomach	1.0%	8.49	5.96	1.51
Bladder	1.0%	5.94	1.67	7.37
Others	0.1%[a]	10.38	4.93	2.39
Total		7.41	4.27	1.96
Female				
Breast	3.6%	4.09	2.51	2.05
Colon	1.5%	10.46	3.95	3.21
Others	0.3%[a]	6.27	3.32	2.27
Total		5.29	2.84	2.33
Male + Female				
Total		6.14	3.35	2.19
Total without breast		7.61	4.02	2.19

[a] Average prevalence of all other sites.

the importance of this conclusion, we consider the data of Lichtenstein *et al.* (3) and interpretation of the data therein in greater detail below.

Evidence from the Twin Data of Lichtenstein *et al.* (3)

We reconsidered the twin concordance data reported by Lichtenstein *et al.* (3). For each cancer site and sex, the data were presented as number of concordant (MZ or DZ) affected pairs (which we denote c), the number of discordant (one affected, one not) pairs, which we denote d, and the number of concordant unaffected pairs, which we denote u. Note that $c + d + u = n$, where n is the total number of pairs. The twin relative risk is then given by $4cn/(2c + d)^2$ for each zygosity. Most of the tissue sites have cancer rates that are too low to allow reliable calculation of λ values. Thus, we calculate λ individually for each site with a prevalence K [$= (2c + d)/2n$] of at least 1% but combine the remaining sites. For these remaining sites, we obtain a weighted average λ value by calculating:

$$4n \sum_{i=1}^{I} c_i / \sum_{i=1}^{I} (2c_i + d_i)^2$$

where i indexes a total of I cancer sites. Results are given in Table 3 individually for prostate, lung, colon, stomach, and bladder cancer among men, with all other 20 sites combined, and for breast and colon cancer in women, with all other 24 sites combined. The sites are listed in order of decreasing prevalence. As can be seen in the table, there is no pattern of decreasing values of λ_M and λ_D with decreasing K, nor does the value of R_{MD} decrease with K. Rather, there appear to be relatively constant values of λ_M and λ_D with K. To test this, I performed a goodness-of-fit test of a model with fixed values for λ_M and λ_D (6.14 and 3.35, respectively) for both men and women using a likelihood ratio goodness-of-fit χ^2 test for each site and also for all sites combined within four sex-zygosity groups: MZ male, DZ male, MZ female, and DZ female. The model fit poorly, because three zygosity-site combinations in males gave significant values (MZ larynx, $\chi^2 = 7.48$; MZ prostate, $\chi^2 = 3.87$; DZ lung, $\chi^2 = 8.38$) and four combinations in females were significant (MZ stomach, $\chi^2 = 3.91$; MZ colon, $\chi^2 = 6.01$; MZ breast, $\chi^2 = 11.82$; DZ breast, $\chi^2 = 5.94$). Also, the overall fit was poor: MZ male, $\chi^2 = 3.56$; DZ male, $\chi^2 = 5.06$; MZ female, $\chi^2 = 2.10$; DZ female, $\chi^2 = 2.58$. Summing these last four gives $\chi^2 = 13.30$ (with four degrees of freedom, $P < 0.01$).

It is apparent that the poor fit is attributable to the discrepancy between female breast cancer (the most common site), which has lower values of λ_M and λ_D than do the other sites. Testing the model of a constant value of λ_M and λ_D for all sites other than female breast gave an excellent fit to the data with $\lambda_M = 7.61$ and $\lambda_D = 4.02$ (results given in Table 4). No individual site-zygosity tests in females were significant, although MZ larynx ($\chi^2 = 6.54$) and DZ lung ($\chi^2 = 4.40$) in males were significant. Considering the total number of site-zygosity combinations ($n = 100$) tested, however, these results should not be considered formally significant. Furthermore, the overall fit to the four sex-zygosity groups was excellent, with the total $\chi^2 = 1.01$ (four degrees of freedom, $P = 0.90$). Thus, a model of constant λ_M (7.61) and λ_D (4.02) for all sites in both sexes, but lower values for breast cancer in women ($\lambda_M = 4.09$, $\lambda_D = 2.51$), gave an excellent overall fit to the data.

This analysis, based on the results in Table 3, allows us to come to the important conclusion that the values of λ_M and λ_D are reasonably consistent across individual cancer sites and do not decrease with decreasing prevalence (K) of the cancer site. One can see, however, that in the context of the MFT model, heritability estimates would decrease with decreasing K because, as was indicated above, the estimate of H does decrease with K for constant λ values. For example, as seen in Table 1, using a value of $\lambda_M = 6.1$, for a cancer with prevalence $K = 0.1\%$, H would be estimated at ~20%, whereas for a cancer with prevalence $K = 3\%$, H would be estimated at ~50%. Similarly, for $\lambda_D = 3.4$, for a cancer with prevalence 0.1%, H is ~23%, whereas for a cancer with prevalence 3%, H is 60%. Thus, the conclusion that rarer cancers are less heritable (3) is strictly a consequence of the assumptions of the MFT model and is not robust to violations of that model. For example, if instead we measure gene effects directly in terms of λ_M and λ_D values, there is no such decrease with prevalence.

Another observation in Table 3 requires mention. The ratio R_{MD} does not decrease systematically with decreasing K, indicating that by this measure also rarer cancers are not less heritable. Furthermore, the value of R_{MD} for all cancer sites hovers ~2.0 (2.05 for breast cancer and 2.19 for all other cancers). As shown in Table 1, for the MFT model, R_{MD} should range from ~2.5 for a common cancer ($K = 3\%$) to ~4.0 (depending on H) for a rare cancer ($K = 0.1\%$). Thus, the observed value of R_{MD} conforms poorly to the predictions of the MFT model but extremely well to the single-locus or additive genetic model described above, which predicts $R_{MD} = 2$. Thus, it is more likely that genetic susceptibility to cancer, in general, entails rare dominant genes and/or additive gene effects across contributing loci than genetic interactions.

It is also important to consider the consequences for genetic analysis based on the MFT model when the actual value of R_{MD} is 2, whereas the MFT model predicts a higher value (*e.g.*, $R_{MD} = 4$). Application of the MFT model often allows for division of the environmental component of liability into a random component and a component attributable to shared twin environment S (3), which is assumed to increase the correlation between MZ and DZ twins to the same extent (as opposed to the genetic component which increases the MZ correlation 2-fold compared with the DZ correlation). The inclusion of such a component thus leads to attenuation of R_{MD} from the value expected for a polygenic model without S. It is therefore predictable that when the single-locus or additive model is correct and $R_{MD} = 2$, application of the MFT model will lead to a positive estimate of S. This would be especially true for rare

Table 4 Goodness-of-fit of constant risk ratio model to cancer concordance data in twins

Sex and site	MZ					DZ				
	Concordant		Discordant		χ^2	Concordant		Discordant		χ^2
	Obs[a]	Exp	Obs	Exp		Obs	Exp	Obs	Exp	
Male										
Prostate	40	37.8	299	303.4	0.20	20	28.4	584	567.2	3.29
Lung	15	18.2	233	226.6	0.78	24	15.7	416	432.6	4.40
Colon	10	13.0	202	196.1	0.92	17	13.3	393	400.4	1.08
Stomach	6	5.4	131	132.2	0.08	8	5.4	256	261.2	1.20
Bladder	5	6.4	146	143.2	0.39	2	4.8	253	247.4	2.25
Others	9	6.5	559	563.8	0.90	7	5.8	999	1001.5	0.24
Total	85	87.3	1570	1565.3	0.08	78	73.4	2901	2910.3	0.32
Female										
Colon	20	14.5	214	224.9	2.37	15	15.3	453	452.5	0.01
Others	16	19.5	1082	1075.2	0.71	18	21.8	2103	2095.4	0.73
Total	36	34.0	1296	1300.1	0.14	33	37.1	2556	2547.9	0.47

[a] Obs, observed; Exp, expected. Concordant unaffected not shown.

cancers, where the observed R_{MD} (= 2) deviates more from the R_{MD} expected for a pure polygenic model. Hence, the conclusion of a significant shared twin environmental component may simply be a consequence of using the wrong genetic model (*i.e.,* MFT *versus* single-locus/additive), rather than indicating the actual existence of such a factor.

It is also of interest to compare the numbers in Tables 2 and 3. For dominant gene effects, the FRR, as given in Table 2, should correspond to λ_D of Table 3. Because in Utah FRR was based on all first-degree relatives, including parents and offspring as well as siblings, FRR might, in theory, be less than λ_D if recessive genes are involved in cancer susceptibility. In fact, the average value of FRR estimated in Table 2 is ~2.12, compared with 3.4–4.0 for λ_D observed in Table 3. At first glance, this might suggest the presence of recessive genes. However, it is important to consider differences in age-structure and follow-up in the various studies.

As opposed to the results of Goldgar *et al.* (7), which were based on age-adjusted lifetime rates, the prevalence figures given in Table 3 do not correspond to lifetime risks. This is because the cohorts of twins were only surveyed for cancer risk during a defined and limited period of time, *i.e.,* they were both left-censored and right-censored (*e.g.,* see Ref. 16). Specifically, Swedish cohort I entered observation between ages 36 and 75 and was followed for 34 years (to ages 70–109, or death); 4,490 subjects of 21,006 (21%) had a diagnosis of cancer. Similarly, the Danish cohort entered study between ages 13 and 73 and were followed for 50 years to ages 63–123 (or death); 3,572 people of 16,922 (21%) had a cancer diagnosis. By contrast, Swedish cohort II entered observation at ages 14–46 and were only followed for 22 years (to ages 36–68); not surprisingly, only 1,157 cancer diagnoses were made in this group of 25,716 subjects studied (*i.e.,* 4.5%). Similarly, the Finnish cohort entered at ages 18–96 and were followed for only 20 years to ages 38–116 or death. There were 1,584 cancer diagnoses of 25,882 subjects (6.1%). Swedish cohort II and the Finnish cohort represent more than half of all of the twin pairs (25,824 of 44,788; 58%). The large majority of cancers in these two twin cohorts have yet to occur.

The values of λ_M and λ_D given in Table 3 are likely to be strongly influenced by the age-structure of the sample. For example, it is known that familiality of many cancers (such as breast cancer) is higher at an earlier age of diagnosis, and thus λ values decrease with age. The numbers provided in Table 3 correspond to cancers diagnosed primarily in midlife. Indeed, Table 2 also provides FRR values for cancers occurring in midlife (before age 50 or 60) for 10 cancer sites. The average value for these early-occurring cancers is 3.8, very close to the average value of λ_D given in Table 3, as well as the sibling recurrence risk ratio from Sweden (Table 2). Thus, it appears most likely that the more modest values of FRR in Table 2 *versus* values of λ_D in Table 3 reflects the different age-structure and follow-up of the twin samples rather than the presence of recessive genes.

Inherited Susceptibility: Site Specific or Generalized?

Another interesting observation from the study of Lichtenstein *et al.* (3) is the MZ and DZ concordances when all cancer sites were grouped together and analyzed as a single entity. For example, such an analysis considers pairs concordant if one twin has lung cancer and the other colon cancer. For male MZ pairs, 262 were considered concordant *versus* 1,252 discordant, giving λ_M = 2.40. For male DZ pairs, 356 were concordant and 2,459 discordant, or λ_D = 1.95; also, R_{MD} = 1.47. For female MZ pairs, 265 were concordant and 1,487 discordant, or λ_M = 2.2, whereas for female DZ pairs, 408 were concordant and 3,023 discordant giving λ_D = 1.70, and R_{MD} = 1.7. These values of λ and R_{MD} are considerably attenuated from the corresponding numbers calculated from site-specific analyses. This observation indicates that the inherited predisposition to cancer is likely to involve many genes that are primarily (but not entirely) site specific. Goldgar *et al.* (7) also examined all pairs of cancer sites in their probands and affected first-degree relatives to ascertain possible genetic relatedness of susceptibility to different cancer sites. This involved consideration of 1,026 comparisons. Despite the large number of tests, many were deemed to be statistically significant. However, from a global perspective, considering all site pairs, the FRR was considerably reduced compared with "within site" estimates, consistent with the observations from the twin data. Also, the analysis of the Swedish family study considered across-site comparisons (8). Although many of these comparisons were statistically significant, they also generally found that the highest risk ratios were associated with site-specific recurrence. The conclusion of site specificity is also consistent with molecular results that to date have shown gene effects to be largely site specific (*e.g.,* colon cancer and melanoma) and/or with very limited range (*e.g.,* breast/ovarian cancer).

Table 5 Penetrances (f_0, f_1), RR_{Het}, and PAF for the single-locus (additive) genetic model as a function of disease allele frequency p, assuming prevalence $K = 1\%$

λ_M (λ_D)	$P = 0.001$	$P = 0.005$	$P = 0.01$	$P = 0.05$	$P = 0.10$
4.0 (2.5)					
PAF	7.7%	17.4%	24.6%	56.2%	81.6%
RR_{Het}	43.0	22.0	17.3	13.8	23.2
f_0	0.0092	0.0083	0.0075	0.0044	0.0018
f_1	0.396	0.182	0.131	0.0605	0.0426
7.4 (4.2)					
PAF	11.3%	25.4%	36.0%	82.1%	
RR_{Het}	64.8	35.0	29.1	46.8	
f_0	0.0089	0.0075	0.0064	0.0018	
f_1	0.576	0.261	0.186	0.084	

Heritability *versus* Attributable Risk

Heritability, as defined above in the context of the MFT model, is often used as a measure of the importance of genetic effects. For example, in the study of Lichtenstein *et al.* (3), because the heritabilities were generally estimated at <30%, the authors concluded that genetic effects are minor relative to environmental impacts. However, when the MFT model does not apply, it is not clear what such heritability estimates indicate. To quantitate the impact of risk factors, epidemiologists use several other measures, most notably the relative risk RR (risk to exposed *versus* unexposed individuals) and *PAF* (proportion of disease prevented by elimination from the population of the risk factor). I now show that the derived twin relative risks (λ_M and λ_D as given in Table 3) are compatible with a broad range of *RR* and *PAF* values for individual gene effects, including very high ones (see the "Appendix" for details of calculations).

For two values of λ_M (and hence λ_D) and allele frequencies ranging from 0.001 to 0.10, the values of *PAF* and RR_{Het} have been calculated (Table 5). The λ values correspond approximately to what was observed for breast cancer and all other cancers combined. Also, the values of f_1 and f_0 corresponding to $K = 1\%$ are given. Comparable values of f_1 and f_0 can be obtained for other values of K simply by multiplication. For example, for $K = 0.1\%$, the values of f_1 and f_0 in Table 5 are divided by 10.

Table 5 demonstrates that the PAF can range from small values to 100%, depending on the disease allele frequency. We note that the disease allele here represents the sum total of all susceptibility alleles at one or multiple loci (*e.g.*, a total frequency p of 1% may correspond to 10 alleles each with frequency 0.1%). Thus, if susceptibility alleles are common, the reported twin data are consistent with up to 100% of cancer being attributable to inherited susceptibility. Of course, this would not exclude environmental factors as also important but rather that genetic predisposition is a necessary but not sufficient cause and depends on environmental interactions. On an individual level, the high values of RR_{Het} also demonstrate that the risks can be substantially increased in gene carriers *versus* noncarriers.

For common cancers (such as breast cancer in women), an allele frequency $p = 0.001$ is not consistent with the data because it would imply a heterozygote penetrance >1 (*e.g.*, if the prevalence of breast cancer is 3.6%, the corresponding value of f_1 would be $3.6 \times 0.396 = 1.43$). For breast cancer, studies of *BRCA1* and *BRCA2* alone already suggest a higher total allele frequency than $p = 0.001$ and a higher *PAF* but lower RR_{Het}.

Conclusions: Implications of Family and Twin Studies for Molecular Genetic Research

Just as heritability estimates from the MFT model applied to family and twin studies may not be the optimal measure of genetic impact when the MFT model does not apply, they should also not necessarily be viewed as a good predictor of the ease with which molecular genetic analysis can identify the actual susceptibility genes involved. In fact, looking historically, one would draw the conclusion that molecular genetic success is either independent of or negatively correlated with estimated heritability from twin studies. For example, there has been considerable success for breast (two genes) and colon cancer (four genes), diseases which appear to have relatively modest heritabilities (3). By contrast, success has been much more limited for diseases with high MZ:DZ concordance ratios and heritability estimates, such as schizophrenia, multiple sclerosis, and autism. One plausible explanation is that the MFT model as used in heritability estimation, with many common interacting genes, provides a closer representation to reality for the latter group of disease than for cancer, where susceptibility may be attributable to fewer, rarer genes with little interaction.

Most of the cancer sites listed in Table 2 have a FRR close to 2, whereas most sites also have stable values for λ_M and λ_D and R_{MD} as given in Table 3, although for the rarer cancers reliable values of λ_M and λ_D are not possible. From Table 2, a few sites have particularly low or high values of FRR. For example, uterine and pancreatic cancer and Hodgkin's lymphoma all have FRR values less than 1.32. From the twin data (3), uterine cancer has a λ_M of 2.2 and λ_D of 4.7. Both of these values are greater than the FRR of 1.32 (Table 2), but the higher λ_D than λ_M is also not consistent with a role of genetic susceptibility. For pancreatic cancer, for males and females combined, $\lambda_M = 11.0$ and $\lambda_D = 1.7$. The low value of λ_D is comparable with the observed FRR value of 1.25, but the higher λ_M value is suggestive of genetic susceptibility, albeit perhaps multigenic. For Hodgkin's lymphoma (FRR = 1.25), too few cases were observed in the large population-based twin study (3) for meaningful analysis. However, a prior clinic-based study (13) found 0 of 187 = 0% of DZ twins *versus* 10 of 179 = 5.6% of MZ twins to be concordant. The very high observed λ_M in that study is again suggestive of genetic susceptibility but possibly recessive and/or multigenic.

For the sites in Table 2 with high FRR values (thyroid, multiple myeloma, leukemia, larynx, and testis), most were too rare to obtain individual λ_M and λ_D estimates from the large twin study (3). However, combining across all five of these sites, weighted averages of $\lambda_M = 17.1$ and $\lambda_D = 5.6$ are obtained. These values are higher than the weighted average for all sites combined (even excluding breast) as given in Table 3, suggesting that for these rare sites genetic influence may be more prominent, consistent with the FRR results in Table 2. The higher value of R_{MD} (=3.5) in this case, however, may indicate recessive and/or multiple interacting genes.

What are the implications of these results for molecular strategies to identify the genes underlying cancer susceptibility? To some extent, the answer depends on how many genes are involved in susceptibility to cancer of a specific site, their frequency, penetrance, and interactions. According to the analysis provided above, the data are generally most consistent with either rare dominant alleles or additive gene effects. For rare dominant alleles, the best approach is linkage analysis with multiplex pedigrees. Even if different mutations are involved in different families, there will still be adequate power in this approach provided the heterogeneity is mostly allelic (small number of loci) rather than nonallelic (large number of loci). Indeed, this approach proved successful for both breast cancer (17, 18) and colon cancer (19, 20), despite the fact that several distinct loci were involved. Furthermore, the observation in Table 2 that early age at diagnosis appears to be generally associated with increased familial risk argues for special priority given to families with young ages at diagnosis.

Although this strategy has worked for finding breast and colon cancer genes, to date it has been less successful in leading to clearly replicable linkage results for prostate cancer (21, 22). It is interesting to reexamine Table 3 in this regard. For breast and colon (sex-averaged) cancers, R_{MD} is very close to 2.0. For prostate cancer, R_{MD} was estimated at 3.86. A previous, comparably sized twin study of prostate cancer (9) found an MZ concordance of 27.1% and DZ concordance of 7.1% *versus* a prevalence of 3.17%. These rates translate into values of $\lambda_{\mathrm{M}} = 8.55$, $\lambda_{\mathrm{D}} = 2.24$, and $R_{\mathrm{MD}} = 6.09$. The λ_{M} and λ_{D} values for both studies are quite similar, and the R_{MD} value appears to be significantly >2.0. Thus, it may turn out for this cancer site that the genetic basis is not explained by independent, rare, autosomal dominant mutations but rather by recessive and/or multiple interacting loci. If such is the case, it would be more difficult to obtain a clear linkage signal than was true for breast and colon cancer.

Linkage analysis also requires the penetrance (probability for gene carriers to become affected) to be moderate to high, as otherwise extended multiplex families will not occur. Under these circumstances, nonallelic heterogeneity could become a more serious problem, because individual families will only provide modest LOD scores, and statistical significance would require lumping together many small families. In this case, a practical option is to study genetic isolates or founder populations, because these populations are likely to have considerably reduced allelic heterogeneity compared with outbred populations. Examples of such populations are Mennonites, Ashkenazi Jews, French Canadians, and Finns. Indeed, in the Ashkenazi Jewish population, only two common and one rare mutation occur at *BRCA1* and *BRCA2* loci, as compared with other more outbred populations that have much greater allelic diversity (23). In addition, these founder mutations are typically of relatively recent origin and thus show linkage disequilibrium (allelic association) up to a substantial genetic distance along the chromosome (perhaps a megabase or more), aiding gene identification.

What if the susceptibility alleles are common and panethnic? In this case, candidate gene studies using case-control methods are likely to be more fruitful (24); however, even in this scenario multiplex families are likely to provide greater power than singleton cases (25).

What about cancers which have an identified, major environmental component such as lung cancer and cigarette smoking? According to Table 2, lung cancer appears to be familial (FRR = 1.7–3.2), but the twin data provide nearly equal values of λ_{M} (6.27) and λ_{D} (6.14) in males. The latter would suggest a strong environmental effect shared by twins (*i.e.,* smoking behavior) rather than a genetic component. Ironically, twin studies have consistently shown greater concordance for smoking behavior in MZ twins than DZ twins. This clearly is an example of an environmental exposure being confounded with genetic influence in a twin study paradigm. Yet, paradoxically, this concordance difference in smoking behavior is not reflected in a concordance difference for lung cancer. A comparable study of United States male twins (7) found the same thing–greater concordance in smoking for MZ *versus* DZ twins, yet no difference in concordance for lung cancer. On the other hand, lung cancer in female twins (3), where the prevalence is much lower, does appear to follow a more genetic pattern, $\lambda_{\mathrm{M}} = 21.3$ and $\lambda_{\mathrm{D}} = 1.76$, although these figures are based on small numbers.

When a major environmental exposure is involved in cancer susceptibility, the question becomes: Are there specific genes that increase the risk of cancer in exposed individuals? In unexposed individuals? And are these genes the same? In theory, family and twin studies can address these questions. For example, if the genes are the same, then the risk of cancer should be increased in family members who are both exposed and unexposed, when the index subject is exposed or unexposed. Different genetic mechanisms would imply that only exposed family members of exposed probands are at increased risk. Although the numbers are small, the lung cancer twin data for females *versus* males is suggestive of more pronounced genetic influence on unexposed or less exposed individuals.

In conclusion, taken at face value, family and twin data support a comparable genetic component for susceptibility to most cancer sites, including the rarer or "sporadic" ones. However, family and twin studies have little power to disentangle interactions between unmeasured genes and environmental risk factors or to eliminate confounding between genes and environmental effects that are correlated in relatives. These limitations preclude the possibility of strong inferences about the genetic input to the various forms of cancer. However, studies of spouses of cancer cases show little in terms of increased risks for these unrelated but cohabiting individuals (26), suggesting that environmental factors may indeed contribute little to familial aggregation for most cancers. Thus, the empirical data along with molecular genetic results for at least a few cancers (*e.g.,* breast and colon) support the continued search for cancer susceptibility genes for all cancer sites, in addition to the environmental risk factors with which they interact.

Appendix

Calculation of PAF *and RR to Heterozygotes* (RR$_{Het}$) *for a Single-Locus Model*

Consider a single-locus model with two alleles *S* and *s* with frequencies *p* and $q = 1 - p$, respectively, where *S* is associated with increased susceptibility. Suppose the risk to be affected for individuals with genotype *ss* is f_0, whereas for those with genotype *Ss* it is $f_1 = f_0 + a$ and for genotype *SS* it is $f_2 = f_0 + 2a$ (*i.e.,* an additive model). According to this model, λ_{M} is given by $1 + w$ and λ_{D} by $1 + w/2$, where $w = 2pqa^2/K^2$, and *K*, the population prevalence, is given by $K = f_0 + 2pa$ (12). The heterozygote RR (RR_{Het}) for genotype *Ss versus ss* is given by $f_1/f_0 = 1 + a/f_0$ and the *PAF* for locus *S* is given by $2pa/(f_0 + 2pa)$. Note that $w = 2pqa^2/(f_0 + 2pa)^2$, so that $q/(f_0 + 2pa) = (w/2pq)^{1/2}$ and $PAF = 2pa/(f_0 + 2pa) = (2pw/q)^{1/2}$. Simple algebra shows that $f_0 = K(1 - PAF)$, so that $RR_{\mathrm{Het}} = f_1/f_0 = 1 + a/[K(1 - PAF)] = [2pq + (q - p)(2pqw)^{1/2}/[2pq - 2p(2pqw)^{1/2}$. Thus, although the values of f_1 and f_0 depend on *K*, the values of *PAF* and RR_{Het} depend only on *p* and w ($= \lambda_{\mathrm{M}} - 1$) and not on *K*.

References

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* for the International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature (Lond.) *409:* 860–921, 2001.

2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* The sequence of the human genome. Science (Wash. DC), *291:* 1304–1351, 2001.

3. Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Xkytthe, Z., and Hemminki, K. Environmental and heritable factors in the causation of cancer. N. Engl. J. Med., *343:* 78–85, 2000.

4. Hoover, R. Cancer–Nature, nurture or both. N. Engl. J. Med., *343:* 135–136, 2000.

5. Khoury, M. J., Beaty, T. H., and Cohen, B. H. Fundamentals of Genetic Epidemiology. New York: Oxford University Press, 1993.

6. Risch, N. Linkage strategies for genetically complex traits. I. Multi-locus models. Am. J. Hum. Genet., *46:* 222–228, 1990.

7. Goldgar, D. E., Easton, D. F., Cannon-Albright, L. A., and Skolnick, M. H. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J. Natl. Cancer Inst., *86:* 1600–1608, 1994.

8. Dong, C., and Hemminki, K. Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. Int. J. Cancer, *92:* 144–150, 2001.

9. Sorensen, T. I. A., Nielsen, G. G., Andersen, K. A., and Teasdale, T. W. Genetic and environmental influences on premature death in adult adoptees. N. Engl. J. Med., *318:* 727–732, 1988.

10. Braun, M. M., Caporaso, N. E., Page, W. F., and Hoover, R. N. Genetic component of lung cancer: cohort study of twins. Lancet, *344:* 440–443, 1994.

11. Braun, M. M., Caporaso, N. E., Page, W. F., and Hoover, R. N. A cohort study of twins and cancer. Cancer Epidemiol. Biomark. Prev., *4:* 469–473, 1995.

12. Page, W. F., Braun, M. M., Partin, A. W., Caporaso, N., and Walsh, P. Heredity and prostate cancer: a study of World War II veteran twins. Prostate, *33:* 240–245, 1997.

13. Mack, T. M., Cozen, W., Shibata, D. K., Weiss, L. M., Nathwani, B. N., Hernandez, A. M., Taylor, C. R., Hamilton, A. S., Deapen, D. M., and Rappaport, E. B. Concordance for Hodgkin's disease in identical twins suggesting genetic susceptibility to the young-adult form of the disease. N. Engl. J. Med., *332:* 413–418, 1995.

14. Ahlbom, A., Lichtenstein, P., Malmstrom, H., Feychting, M., Hemminki, K., and Pedersen, N. L. Cancer in twins: genetic and nongenetic familial risk factors. J. Natl. Cancer Inst., *89:* 287–293, 1997.

15. Verkasalo, P. K., Kaprio, J., Koskenvuo, M., and Pukkala, E. Genetic predisposition, environment and cancer incidence: a nationwide twin study in Finland, 1976–1995. Int. J. Cancer, *83:* 743–749, 1999.

16. Marenberg, M. E., Risch, N., Berkman, L. F., Floderus, B., and de Faire, U. Genetic susceptibility to death from coronary heart disease in a study of twins. N. Engl. J. Med., *330:* 1041–1046, 1994.

17. Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., and King, M. C. Linkage of early-onset familial breast cancer to chromosome 17q21. Science (Wash. DC), *250:* 1684–1689, 1990.

18. Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., Averill, D., *et al.* Localization of a breast cancer susceptibility gene, *BRCA2,* to chromosome 13q12–13. Science (Wash. DC), *265:* 2088–2090, 1994.

19. Peltomaki, P., Aaltonen, L. A., Sistonen, P., Pylkkanen, L., Mecklin, J. P., Jarvinen, H., Green, J. S., Jass, J. R., Weber, J. L., Leach, F. S., *et al.* Genetic mapping of a locus predisposing to human colorectal cancer. Science (Wash. DC), *260:* 810–812, 1993.

20. Lindblom, A., Tannergard, P., Werelius, B., and Nordenskjold, M. Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. Nat. Genet., *5:* 279–282, 1993.

21. Smith, J. R., Freije, D., Carpten, J. D., Gronberg, H., Xu, J., Isaacs, S. D., Brownstein, M. J., Bova, G. S., Guo, H., Bujnovszky, P., Nusskem, D. R., Damber, J. E., Bergh, A., Emanuelsson, M., Kallioniemi, O. P., Walker-Daniels, J., Bailey-Wilson, J. E., Beaty, T. H., Meyers, D. A., Walsh, P. C., Collins, F. S., Trent, J. M., and Isaacs, W. B. Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. Science (Wash. DC), *274:* 1371–1374, 1996.

22. Xu, J., and the International Consortium for Prostate Cancer Genetics. Combined analysis of hereditary prostate cancer linkage to 1q24–25: results from 772 hereditary prostate cancer families from the international consortium for prostate cancer genetics. Am. J. Hum. Genet., *66:* 945–957, 2000.

23. Szabo, C. I., and King, M. C. Population genetics of *BRCA1* and *BRCA2*. Am. J. Hum. Genet., *60:* 1013–1020, 1997.

24. Risch, N., and Merikangas, K. The future of genetic studies of complex human diseases. Science (Wash. DC), *273:* 1516–1517, 1996.

25. Risch, N., and Teng, J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res., *8:* 1273–1288, 1998.

26. Hemminki, K., and Dong, C. Life style and cancer: protection from a cancer-free spouse. Int. J. Cancer, *87:* 308–309, 2000.

Cancer Epidemiology, Biomarkers & Prevention

American Association for Cancer Research

The Genetic Epidemiology of Cancer: Interpreting Family and Twin Studies and Their Implications for Molecular Genetic Approaches

Neil Risch

Cancer Epidemiol Biomarkers Prev 2001;10:733-741.