Carcinogenesis vol.32 no.9 pp.1295–1298, 2011
doi:10.1093/carcin/bgr087
Advance Access publication May 11, 2011

REVIEW

Genome-wide association studies may be misinterpreted: genes versus heritability

Paolo Vineis[1,2,*] and Neil E.Pearce[3]

[1]Molecular and Genetic Epidemiology Unit-HuGeF, MRC/HPA Centre for Environment and Health, School of Public Health, Imperial College, Norfolk Place, W2 1PG London, UK, [2]Molecular and Genetic Epidemiology Unit, HuGeF Foundation, Via Nizza 52,—10126 Torino, Italy and [3]Department of Medical Statistics, Faculty of Epidemiology and Public Health, London School of Hygiene and Tropical Medicine, WC1E 7HT London, UK

*To whom correspondence should be addressed. Tel: +44 20 75943372;
Fax: +44 20 75943196;
Email: p.vineis@imperial.ac.uk

Much of the literature on genome-wide association studies (GWAS) is based on the premise that an important proportion of common diseases is *heritable* and that this proportion is likely to be due to genetic variants detectable with extensive scans of the DNA. Heritability is estimated from family studies, including twin studies and is based on the comparison of the variation in disease among different members of particular families. Since there is a wide gap between the population variation in disease explained by the results of GWAS (usually <10% for common diseases) and estimates of heritability (often >50%), the question arises as to how to explain these differences. However, the premise for this question is based on two sources of misunderstanding: (i) confusion between variation and causation and (ii) confusion between heritability and genetic determination. As we show with a number of examples, variation is not causation and heritability is not genetic determination. Therefore, heritability studies do not provide valid estimates of the proportion of disease cases that are attributable to genetic factors. Such estimates in turn cannot be used to estimate the proportion of cases that are due to environmental factors.

Highly relevant discoveries and clinical applications are expected to stem from a systematic approach to the genetic determinants of disease. After the first draft of the human genome and the development of the HapMap program, we have seen a flourishing of research initiatives driven by high-throughput technologies (1). Much of the literature on Genome-Wide Association Studies (GWAS) is based on the premise that an important proportion of common diseases is *heritable*, and that this proportion is likely to be due to genetic variants detectable with extensive scans of DNA. This assumption is crucial for the interpretation of GWAS results, and it is frequently used to justify expectations from researchers (epidemiologists and geneticists), the public at large, funders and the drug industry. For example, this premise is clearly stated in a recent paper by Manolio *et al* (1), where they list a number of diseases, encompassing Type-2 diabetes, height, high-density lipoproteins, fasting glucose and others, for which the proportion of heritability currently explained by the loci detected by GWAS would be between 1.5 (fasting glucose) and 50% (age-related macular degeneration), implying that other relevant genetic loci remain to be detected.

Since there is a wide gap between the population variation in disease explained by the results of GWAS (usually <10% for common diseases) and estimates of heritability (often >50%), the question arises of how to explain these differences. We believe, however, that the premise itself needs to be challenged. Heritability is estimated from family studies, including twin studies and is based on comparisons of the variation in disease among different members of particular families. Parent–offspring studies are used to compute additive variance, whereas twin studies also measure dominance variance, i.e. the effect of the heterozygous genotype and allele interaction within a single locus.

The example of ankylosing spondylitis

These problems are illustrated by a classic paper on the heredity of ankylosing spondylitis (2), which estimated that 97% of the variation was genetic and 3% environmental. Variance was partitioned into four factors, i.e.: the additive effect of alleles at the same locus (A), the effect of 'shared' environmental factors (C), the interactions between alleles at the same locus (D) and a 'random' environmental component (E). By definition, $A + C + D + E$ is equal to one, i.e. this is a completely additive effect, and gene–environment and gene–gene interactions were not considered separately. The authors used data from two different studies to estimate three of the four factors (2). Additive genetic effects were thus estimated to contribute 97% of the population variance, but a very small share of this was explained by HLA variants. The role of the environment was estimated to be only 3% and this was considered to be due probably to 'something ubiquitous' (note that by 'population' variance the authors in fact mean variation across twin pairs). There was therefore a large gap between the heritability estimate and the estimate of the contribution of genetic effects, a gap that—in the authors' opinion—must be due to non-HLA genes. We consider, however, that such interpretations of data on heritability are questionable because they are based on two sources of misunderstanding: (i) confusion between variation and causation and (ii) confusion between heritability and genetic determination.

A source of misunderstanding: variation and causation

The difference between variation and causation is crucial. In particular, the percentage of population variation in a disease due to a particular exposure or trait should not be confused with the proportion of disease explained by this exposure or trait. Let us consider phenylketonuria (PKU) (3). It can be avoided either by avoiding transmission of the PKU gene mutations to the offspring or by a dietary intervention (low phenylalanine in diet). In this sense, PKU cases in the population are 100% attributable to the mutation since 100% of cases would be prevented if it were possible to remove the mutation from the population; however, 100% of cases would also be prevented if everyone adopted a low-phenylalanine diet. If we study a population where everyone has a high-phenylalanine diet, but only some people have the mutation, then 100% of the population variation will be due to the variation in the genetic mutation and 0% will be due to diet (i.e. the condition will appear to be 100% genetic). In contrast, if we study a population where everyone has the mutation, but there is variation in diet, then 100% of the population variation will be due to variation in diet and 0% will be due to variation in the mutation (i.e. the condition will appear to be 100% environmental). The problem with this reasoning lies in the fact that it attempts to separate two aspects (gene and environment) that in this specific case—and probably many others—are inseparable. In particular, the usual calculation of heritability attempts to partition the population variation into separate components that add up to (at most) 100% (the same approach was used in the paper on ankylosing spondylitis). However, when we consider causation, rather than variation, there is no requirement for the attributable fractions for each risk factor (genetic, environmental) to sum to 100% (4)—in fact, as we learn more about a particular disease, it is inevitable that the attributable proportions for each risk factor will sum to >100%. In the PKU example, the disease is 100% attributable to the mutation (i.e. 100% of cases could be prevented by removing

Abbreviations: GWAS, genome-wide association study; PKU, phenylketonuria.

the mutation) and 100% attributable to diet (i.e. 100% of cases could be prevented by changes in diet).

More generally, if we completely eliminate genetic variation such as in inbred experimental animals exposed to a strong carcinogen, we obtain a purely 'environmental' causal pattern. If, on the other hand, we choose a population of equally exposed subjects, 100% of the variation will be due to genetic factors. One example of the former 'experiment' was the observation of smoking patterns in twins. When monozygotic twins were studied for both their smoking habits and lung cancer, the latter had a much higher occurrence in the twin who smoked, when pairs were discordant for smoking habits and it occurred with the same frequency when both twins smoked, or when neither smoked, respectively (2,4). Since the genetic variation was zero, all of the variation was due to the 'environment'. Another example is provided in Table I, which shows that the risk of developing tuberculosis is much higher in strict relatives of cases, with a dose–response relationship with the degree of genetic relatedness. The fact that twins show 80% concordance for the development of tuberculosis means that their genetic identity 'amplifies' the effect of the shared environmental exposure, it does not mean that 80% of tuberculosis is 'heritable' or genetically determined 'in the population'. In this specific case of total genetic identity and of shared environmental exposure, concordance between twins is allowed to approach 100%.

The crucial distinction between the analysis of variance and the analysis of causes was clearly established in Lewontin's seminal paper (6) published in 1974. It also follows directly from Rothman's theory of causes in epidemiology (7) in that cases of disease that are attributable to a particular sufficient causal constellation can be prevented by removing any component of the constellation. Despite this, the genetic epidemiology literature is still full of studies that attempt to partition the 'causes' of a particular disease into components that sum to 100%. In particular, it is frequently the case (in fact, it is rarely not the case), that authors estimate that the causation of a particular disease is xx% due to genetic factors and therefore must be 100−xx% due to environmental factors. This fallacious reasoning is based on: (i) assessing the population variation due to genetic factors and then assuming that the estimates obtained are valid estimates of the percentage of cases of disease that are attributable to such factors in the population and (ii) then subtracting this estimate from 100% to estimate the percentage of cases that is due to environmental factors.

Heritability and genetic determination

This problem is compounded by a second source of confusion regarding the difference between heritability and genetic determination. As noted above, the proportion of population variation due to genetic factors is not a valid estimate of the proportion of cases attributable to genetic factors. Furthermore, the proportion of population variation that is due to heritability is not a valid estimate of the proportion that is due to genetic factors.

It should be noted that a major debate on heritability has occurred previously, after the publication of 'The Bell Curve' by Herrnstein and Murray (1994). As it was pointed out (8), the basic confusion in this book and in many other similar papers was between heritability and genetic determination. Heritability involves similar patterns of observable traits between parents and the offspring, while a characteristic is 'genetically determined' if it is coded in and caused by the genes in a 'normal' environment. One extreme example is the following: wearing skirts among European populations has a very strong heritability (it occurs only in women, with the exception of the odd Scotsman and Pacific Islander); it is thus related to having XX versus XY. However, it is not genetically determined (8). A similar example has been provided by Van Asselt *et al.* (9): 'It is crucial to realize that heritability is a ratio, a relative measure. A more uniform environment will increase the heritability, even if the variance in disease occurrence resulting from genetic factors remains the same. A clear illustration, given by Hirsch in 1981, relates to the number of legs that humans have, of which all the variation is determined by environment (amputations, thalidomide). Although the number of legs is determined by genes, because of the absence of genetic variation between humans, the heritability estimate is 0%' (9).

Table I. Effect of genetic relatedness on host response to *Mycobacterium tuberculosis* in families with an index case (5)

Relation of family member to index case	% of exposed and susceptibles showing clinical manifestations of TB
Marriage partner	7.1
Half sibling	11.9
Dizygotic twin	25.5
Monozygotic twin	83.3

TB, tuberculosis.

Such misconceptions are clearly relevant to the discussion about the heritability versus genetic determination of disease. In practice, studies of twins or siblings cannot be used to infer directly that cancer or schizophrenia are due to inherited changes in DNA. Manolio *et al.* (1) make the same error when they list a number of diseases, for which the proportion of heritability currently explained by the loci detected by GWAS is low and conclude that other relevant genetic loci remain to be detected. This reasoning is faulty. One can in fact argue that the environment itself is inherited, for example in the case of the propensity to wear skirts, so the heritability of a disease includes both genetic and environmental factors. For example, claims that IQ has 60% heritability, academic performance 50% and occupational status 40% (8) do not mean that such characteristics are inherited through genes but only that there is a strong association between the characteristic in the index subject and the same characteristic in the parents (the same applies in fact to voting behaviors and religious beliefs).

As stated by Van Asselt *et al.* (9), 'currently, the role of heritability in human studies may be better considered qualitatively with a judgment on whether detectable genetic variance is present and not its magnitude. To accept a quantitative heritability estimate from any study as a fact of nature is but an illusion'.

Other arguments against a strong genetic predisposition to common diseases

Other evidence against a strong genetic predisposition is provided by studies in migrants, which are clearly in favor of a predominant role of the environment in chronic diseases. The Japanese in the 1970s had an annual incidence of stomach cancer that was 133/100 000/year, while it was ∼27 among Americans living in the Hawaii, i.e. five times lower. The Japanese who migrated to the Hawaii had—at the first generation—an incidence of 40/100 000. The incidence further decreased in the Japanese who settled in the USA and reached the incidence of white Americans in the next generation (10). Figure 1 illustrates the same point with data on melanoma in migrants. Studying migrants involves maximizing the variation in exposures and reducing the genetic variation (though not so radically as in twin studies).

Time trends also provide evidence against a strong genetic predisposition to disease. For example, colon cancer was very low in the 1970s in Japan (8/100 000/year, versus 37 in the USA), but has increased markedly in recent decades, reaching the highest incidence in the world (up to 59/100 000/year in men) (12). One of the most striking changes in disease occurrence in recent decades was the rapid decrease of cardiovascular disease incidence in most Western countries (Figure 2). The causes are still uncertain, but it is probably that this can be ascribed to public health measures. In any case, it is very difficult to interpret these figures as suggesting a strong role for genetics. A similar reasoning applies to recent increases of obesity or diabetes.

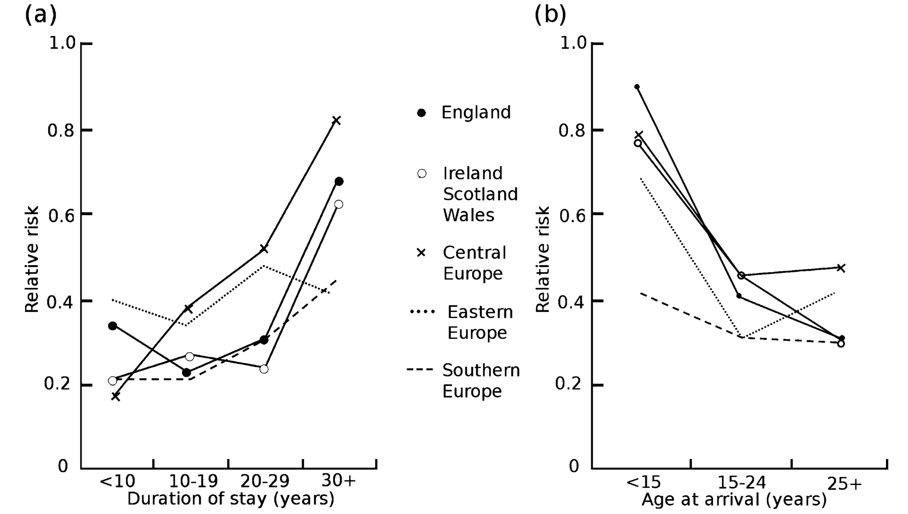

Fig. 1. Relative risk of melanoma as a function of age at migration or duration of stay from low to high-risk areas (14).

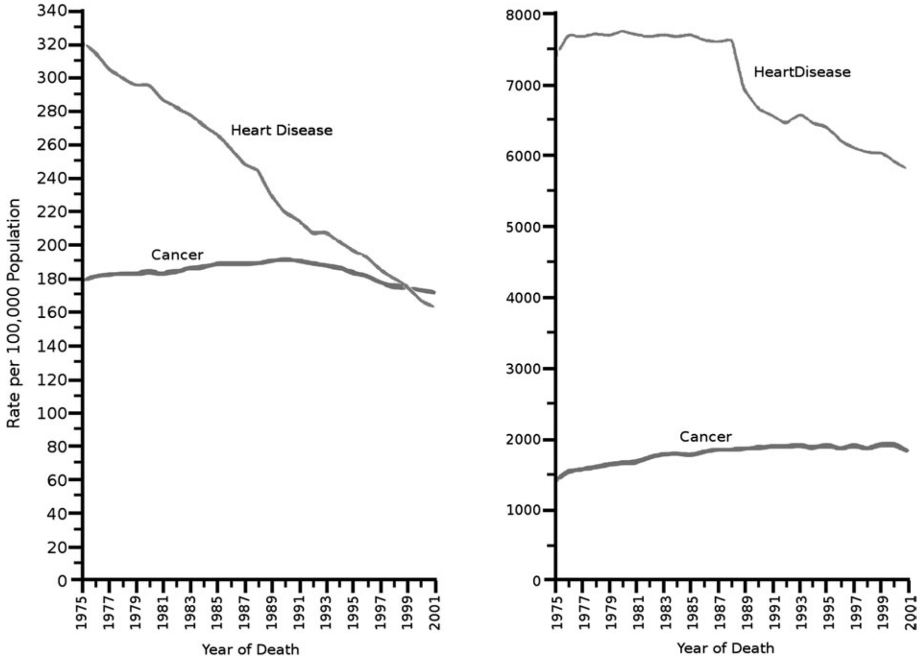

Fig. 2. Death rates (per 100 000/year) from heart disease and cancer (overall) from 1975 to 2001 in the USA. From CDC website: http://www.cdc.gov/nchs/data/hus/hus09.pdf.

The genetic profile in the population and the effect of multiple genetic variants

Most individuals have low-penetrance variants rather than highly penetrant, rare variants and the overall risk will be related more to the combination of several or many variants than to anyone of them in isolation (13). This is reminiscent of the conundrum indicated by Geoffrey Rose years ago in a seminal paper (14): from a public health point of view it is usually not worth focusing on high-risk subgroups (such as highly penetrant mutations, which are rare), because the vast majority of the cases of disease occur in subjects at average or low risk. In fact, Rose's reasoning can be interpreted as having extreme consequences for genetics. One can hypothesize that no two persons will be equal to each other in terms of genetic profile, not only in general but also in relation to specific disease risks. In other words, it will be very difficult to prove that the profiles that are found only among diseased cases indicate causal mechanisms since it is extremely unlikely that the same combination of risk factors will be found in more than one person. When combinations of risk factors are 'unique', only a few other persons in the world may have that exact same profile and screening becomes impractical.

If we apply the same criteria used above to both environmental exposures and genetic variants (Table II), then there are two observations that can be drawn. Firstly, if the majority of the population have the relevant exposure and the genetic variant, the sum of attributable risks is by definition >100%, indicating that a large number of cases can be prevented by acting either on exposure or on the gene. Secondly, the attributable proportion depends on the frequency of exposure (environmental or genetic) in a particular population; thus, assuming generically that there is a 'fixed' proportion due to one or the other in a population makes no sense. Even if we accept the false paradigm that the attributable proportions should sum to 100%, the share attributed to genes or to the environment depends on their frequencies in a particular population.

Table II. Attributable risks in the population as a function of relative risk and proportion of exposed subjects, under the assumption of a linear combination of genes and environment

	Exposure (E)	Genetic variant (G)	Attributable risk	
			E	G
RR	10.0	10.0		
Proportion Exposed	90%	90%	81%	81%
	90%	0%	81%	0%
	50%	50%	45%	45%
	80%	20%	72%	18%
RR	10.0	2.0		
Proportion Exposed	90%	90%	81%	45%
	90%	0%	81%	0%
	50%	50%	45%	25%
	80%	20%	72%	10%
	20%	80%	18%	40%

Conclusions

In conclusion, we consider that the conceptual challenges we have proposed need to be clarified. If 'preventive genomic medicine'—as defined by Francis Collins—is to become a reality, we have to be aware of its likely problems and limitations. In particular, common interpretations of data on heritability are questionable because they are based on two sources of misunderstanding: (i) confusion between variation and causation and (ii) confusion between heritability and genetic determination. Variation is not causation and heritability is not genetic determination. Therefore, heritability studies do not provide valid estimates of the proportion of disease cases that are attributable to genetic factors. Such estimates in turn cannot be used to estimate the proportion of cases that are due to environmental factors.

Funding

Compagnia di San Paolo, Torino, to Paolo Vineis (HuGeF Foundation, 2010). The Centre for Public Health Research is supported by a Programme Grant from the Health Research Council of New Zealand.

Acknowledgements

Based on an invited keynote lecture by P.V. at the Rotterdam University, Department of Epidemiology, 23 February 2010.

We thank Jan Vandenbroucke, at the Leiden University Medical Center. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest Statement: None declared.

References

1. Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
2. Carmelli,D. *et al.* (1996) Twenty-four year mortality in World War II US male veteran twins discordant for cigarette smoking. *Int. J. Epidemiol.*, **25**, 554–559.
3. Pearce,N. (2011) Epidemiology in a changing world: variation, causation and ubiquitous risk factors. *Int. J. Epidemiol*, **40**, 503–512.
4. Braun,M.M. *et al.* (1994) Genetic component of lung cancer: cohort study of twins. *Lancet*, **344**, 440–443.
5. Evans,A.S. (1993) *Causation and Disease*. Plenum Press, New York, NY.
6. Lewontin,R.C. (2006) The analysis of variance and the analysis of causes. 1974. *Int. J. Epidemiol.*, **35**, 520–525.
7. Rothman,K.J. *et al.* (2008) *Modern Epidemiology.* Lippincott, Williams and Wilkins, Philadelphia, PA.
8. Block,N. (1995) How heritability misleads about race. *Cognition*, **56**, 99–128.
9. van Asselt,K.M. *et al.* (2006) Role of genetic analyses in cardiology: part II: heritability estimation for gene searching in multifactorial diseases. *Circulation*, **113**, 1136–1139.
10. Haenszel,W. *et al.* (1968) Studies of Japanese migrants. I. Mortality from cancer and other diseases among Japanese in the United States. *J. Natl Cancer Inst.*, **40**, 43–68.
11. Minami,Y. *et al.* (2006) Increase of colon and rectal cancer incidence rates in Japan: trends in incidence rates in Miyagi Prefecture, 1959–1997. *J. Epidemiol.*, **16**, 240–248.
12. Vineis,P. *et al.* (2008) Expectations and challenges stemming from genome-wide association studies. *Mutagenesis*, **23**, 439–444.
13. Rose,G. (1985) Sick individuals and sick populations. *Int. J. Epidemiol.*, **14**, 32–38.
14. McCredie,M. *et al.* (1990) Cancer incidence in European migrants to New South Wales. *Ann. Oncol.*, **1**, 219–225.

Received January 25, 2011; revised April 06, 2011; accepted April 14, 2011