Gene-Environment Dependence Creates Spurious Gene-Environment Interaction

Frank Dudbridge[1,*] and Olivia Fletcher[2,3]

Gene-environment interactions have the potential to shed light on biological processes leading to disease and to improve the accuracy of epidemiological risk models. However, relatively few such interactions have yet been confirmed. In part this is because genetic markers such as tag SNPs are usually studied, rather than the causal variants themselves. Previous work has shown that this leads to substantial loss of power and increased sample size when gene and environment are independent. However, dependence between gene and environment can arise in several ways including mediation, pleiotropy, and confounding, and several examples of gene-environment interaction under gene-environment dependence have recently been published. Here we show that under gene-environment dependence, a statistical interaction can be present between a marker and environment even if there is no interaction between the causal variant and the environment. We give simple conditions under which there is no marker-environment interaction and note that they do not hold in general when there is gene-environment dependence. Furthermore, the gene-environment dependence applies to the causal variant and cannot be assessed from marker data. Gene-gene interactions are susceptible to the same problem if two causal variants are in linkage disequilibrium. In addition to existing concerns about mechanistic interpretations, we suggest further caution in reporting interactions for genetic markers.

There is much interest in discovering interactions between genetic and environmental risk factors for disease, because such interactions can shed light on biological processes leading to disease, identify subjects for whom risk factors are most relevant, and improve the accuracy of epidemiological risk models.[1] Interaction is commonly understood as the modification by one factor of the effect of the other and is assessed statistically by testing for departure from additivity, on an appropriate scale, of the effects of gene and environment. Such a definition can be dependent on modeling assumptions and might not correspond to biological notions of interaction,[2–4] but it is nevertheless useful in general exploratory settings.

To date, relatively few gene-environment interactions have been reported, in contrast with the large number of marginal associations discovered through genome-wide association studies (GWASs). One reason is that there might be relatively few subjects for whom the joint effect of gene and environment is high, so that very large samples are required to detect interactions. Another is that measurement error in either gene or environment can lead to substantial increases in required sample size.[5] Thus, robust identification of a gene-environment interaction is regarded as a noteworthy finding.

In most studies, genotypes are measured not for the variants that directly affect disease but for markers in linkage disequilibrium (LD) with the causal variants. This is especially true in GWASs and other large-scale discovery studies that aim to map novel disease variants. This creates a misclassification problem in that the true causal variants have been measured with error. In contrast to common measurement error models, a marker is not an unbiased measurement of a causal variant (because the genotype frequencies of the marker and causal variant may differ), and the misclassification probabilities are unknown by design. General methods to adjust for measurement error[6] are not applicable, and so we must accept possible bias in the estimation of interaction effects.

Through simulations, Hein et al.[7] showed that the interaction effect of a marker is biased toward the null, with a corresponding increase in the sample size required for a study based on a marker. Garcia-Closas et al.[8] showed analytically that measurement error in the environmental exposure also biases the interaction effect toward the null. Furthermore, Greenwood et al.[9] showed that the interaction effect is not biased by measurement error in additional covariates included in the model. All of these studies assumed that the genetic marker and environmental exposure are independent in the source population or in the controls. Gene-environment independence also underlies the case-only design[10] and extensions of it designed to improve the power of interaction tests.[11–14] This assumption is often reasonable: for example, autosomal genotypes tend to be independent of sex. However, the properties of interaction tests have not been considered when gene and environment are not independent. Here we demonstrate that under gene-environment dependence, the interaction effect of a marker can be nonzero even if there is no interaction between the causal variant and the environment.

We and others recently established an association with breast cancer of the marker rs10235235, which maps to the *CYP3A* locus.[15] This marker, which was initially

[1]Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK; [2]Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London SW7 3RP, UK; [3]Division of Breast Cancer Research, The Institute of Cancer Research, London SW7 3RP, UK
*Correspondence: frank.dudbridge@lshtm.ac.uk
http://dx.doi.org/10.1016/j.ajhg.2014.07.014.

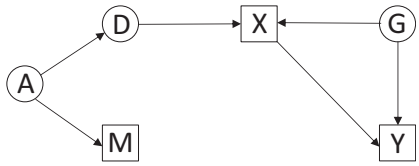

Figure 1. Directed Acyclic Graph Showing Gene-Environment Dependence by Mediation
Boxes are observed variables, circles are unobserved, and arrows indicate directions of causal relationships. Abbreviations are as follows: D, causal variant; M, marker genotype; X, environmental exposure; Y, outcome such as disease; A, composite variable for ancestry, which gives rise to correlation (LD) between D and M; and G, composite variable for common causes of X and Y, which may include additional genes. X mediates the effect of D on Y. For example, a *CYP3A* variant (D) affects the risk of breast cancer (Y) via its effect on age at menarche (X).

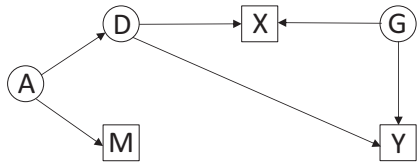

Figure 2. Directed Acyclic Graph Showing Gene-Environment Dependence by Pleiotropy
Notation as in Figure 1. D has pleiotropic effects on X and Y, but there is no direct effect of X on Y. For example, a *CYP3A* variant (D) affects hormone levels, which independently affect age at menarche (X) and breast cancer (Y). Age at menarche is an independent risk factor for breast cancer because it marks additional causal processes (G).

identified through its association with urinary estrone glucuronide,[16] a metabolite that is correlated with the sex hormone estradiol, is also associated with age at menarche. We found a statistical interaction between rs10235235 and age at menarche on breast cancer risk, which is therefore a gene-environment interaction under gene-environment dependence. However, rs10235235 is not known to be the causal variant, and we will show that the marker interaction does not imply an interaction at the causal variant.

Also in breast cancer, Nickels et al.[17] established a statistical interaction between the marker rs3817198 at *LSP1* (MIM 153432) with parity, but also reported significant negative correlation between rs3817198 genotype and number of births. Again, this is a gene-environment interaction in the presence of gene-environment dependence, but rs3817198 is not known to be the causal variant.

As a further example, variants at chromosome 15q25 have been associated with both smoking and lung cancer.[18–20] Interactions between these variants and smoking on lung cancer risk have also been identified,[19,21] but not for the likely causal variants.[22]

Gene-environment dependence could arise in a number of ways. There is likely to be a genetic component to many of the established risk factors for which interactions are sought. In addition to the examples above, variants in genes involved in alcohol metabolism have been associated with alcohol intake,[23] which is a risk factor for many diseases.[24] GWASs have identified numerous variants associated with obesity, an established risk factor for many complex disorders including type 2 diabetes and breast cancer.[25] Similarly, multiple variants that influence low-density lipoprotein cholesterol levels, one of the strongest risk factors for cardiovascular disease, have been identified.[26] Even the more exogenous exposures, such as urban environment, might conceivably have a genetic component.[27] However, on a per-gene basis, knowledge of biological function could be invoked to argue that a given gene is unlikely to affect an exposure of interest.

Figures 1, 2, and 3 illustrate three basic forms of gene-environment dependence. Association of a gene with an environmental risk factor is often taken to imply mediation of the genetic effect by the risk factor. That is, at least part of the effect of the gene on the outcome is via its effect on the environmental factor. For example, a variant at the *CYP3A* locus might directly affect levels of the hormone estradiol, which influences age at menarche, which then directly affects breast cancer risk (Figure 1).

On the other hand, the gene might have pleiotropic effects on the environmental factor and the outcome, but the environment might not cause the outcome. For example, a *CYP3A* variant might, via estradiol levels, influence both age at menarche and breast cancer risk. Age at menarche might have no direct effect on breast cancer, so it does not mediate the effect of the *CYP3A* variant, but it remains a risk factor by acting as a marker for other mechanisms that affect disease (Figure 2).

Gene-environment dependence could also arise through confounding, of which the principal source is population structure. For example, some *CYP3A* haplotypes might have become less frequent at northern latitudes. For unrelated reasons, age at menarche tends to be higher at northern latitudes, leading to an association with *CYP3A* variants (Figure 3). This confounding might be independent of any confounding of the gene and outcome and cannot be corrected using the standard methods to adjust for gene-outcome confounding.

Any combination of the above three forms might occur in practice, so for example a pleiotropic gene might affect the outcome through several pathways, only one of which is mediated by the environment of interest. The corresponding graph would be a combination of Figures 1 and 2, including both direct and indirect effects of the causal variant.

To formalize the interaction effects under gene-environment dependence, let M denote a genotyped marker, coded numerically, X an environmental exposure, and Y an outcome of interest. Consider a generalized linear model

$$E(Y \mid M, X) = h^{-1}(\beta_0 + \beta_M M + \beta_X X + \beta_{MX} MX),$$

where h is a known link function. Writing $\eta_{m,x} = \beta_0 + \beta_M m + \beta_X x + \beta_{MX} mx$, the interaction term in this model is

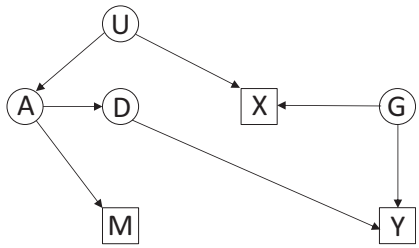

Figure 3. Directed Acyclic Graph Showing Gene-Environment Dependence by Confounding
Notation as in Figure 1, with U an unmeasured confounder. D and X are associated by confounding. For example, haplotype frequencies at CYP3A (A) vary by latitude (U), as does age at menarche (X). In this graph, U is not a confounder of D and Y; such confounders are omitted for simplicity.

$$\beta_{MX} = \eta_{1,1} - \eta_{0,1} - \eta_{1,0} + \eta_{0,0}. \qquad \text{(Equation 1)}$$

Let D denote the unmeasured genotype of the causal variant, with a corresponding generalized linear model in D and X

$$E(Y \mid D, X) = h^{-1}\left(\beta_0^* + \beta_D^* D + \beta_X^* X + \beta_{DX}^* DX\right),$$

where the asterisks denote effects in the model for D rather than for M. If the marker has no effect on the outcome, conditional on D, then

$$E(Y \mid M, X) = \sum_d E(Y \mid d, X)\Pr(d \mid M, X)$$
$$h^{-1}\eta_{m,x} = \sum_d h^{-1}\left(\beta_0^* + \beta_D^* d + \beta_X^* x + \beta_{DX}^* dx\right)\Pr(d \mid m, x). \qquad \text{(Equation 2)}$$

The conditional distribution $\Pr(D|M,X)$ accounts for both the LD between marker and causal variant and the dependence between exposure and causal variant. Equations 1 and 2 allow the marker interaction term to be nonzero even when the interaction term for the causal variant is zero. Some conditions under which the marker interaction term is in fact zero are given in the following lemma.

Lemma

If $\beta_{DX}^* = 0$, then $\beta_{MX} = 0$ if any of the following conditions hold

(1) there is no main effect of the causal variant on the outcome, $\beta_D^* = 0$
(2) the marker is perfectly correlated with the causal variant, $D = M$
(3) the causal variant is independent of the marker, conditional on the exposure, $\Pr(D|M,X) = \Pr(D|X)$

Furthermore, under linear ($h(x) = x$) or log-linear ($h(x) = \log(x)$) regression, $\beta_{MX} = 0$ if

(4) the causal variant is independent of the exposure, conditional on the marker, $\Pr(D|M,X) = \Pr(D|M)$

Proof

If $\beta_{DX}^* = 0$ then the terms in β_{MX} are explicitly

$$h^{-1}(\eta_{1,1}) = \sum_d h^{-1}(\beta_0^* + \beta_D^* d + \beta_X^*)\Pr(d|M=1, X=1)$$
$$h^{-1}(\eta_{1,0}) = \sum_d h^{-1}(\beta_0^* + \beta_D^* d)\Pr(d|M=1, X=0)$$
$$h^{-1}(\eta_{0,1}) = \sum_d h^{-1}(\beta_0^* + \beta_D^* d + \beta_X^*)\Pr(d|M=0, X=1)$$
$$h^{-1}(\eta_{0,0}) = \sum_d h^{-1}(\beta_0^* + \beta_D^* d)\Pr(d|M=0, X=0)$$

If $\beta_D^* = 0$ then

$$h^{-1}(\eta_{1,1}) = \sum_d h^{-1}(\beta_0^* + \beta_X^*)\Pr(d|M=1, X=1)$$
$$= h^{-1}(\beta_0^* + \beta_X^*)$$
$$\eta_{1,1} = \beta_0^* + \beta_X^*$$

Similarly, $\eta_{0,1} = \beta_0^* + \beta_X^*$ and $\eta_{1,0} = \eta_{0,0} = \beta_0^*$, so $\beta_{MX} = 0$ proving (1).

If marker and causal variant are perfectly correlated, then trivially $\beta_{MX} = \beta_{DX}^* = 0$, which proves (2).

If $\Pr(D|M,X) = \Pr(D|X)$, then $\eta_{1,1} = \eta_{0,1}$ and $\eta_{1,0} = \eta_{0,0}$, which proves (3).

Finally, if $\Pr(D|M,X) = \Pr(D|M)$ and either $h(x) = x$ or $h(x) = \log(x)$, then

$$\eta_{1,1} = \beta_0^* + \beta_X^* + \sum_d h^{-1}(\beta_D^* d)\Pr(d|M=1)$$
$$\eta_{1,0} = \beta_0^* + \sum_d h^{-1}(\beta_D^* d)\Pr(d|M=1)$$
$$\eta_{1,1} - \eta_{1,0} = \beta_X^*$$

Similarly, $\eta_{0,1} - \eta_{0,0} = \beta_X^*$ under either link function, so $\beta_{MX} = 0$ as required, which proves (4). Q.E.D.

Conditions (1)–(3) are reassuring because they mean that when a marker-exposure interaction exists, the marker must be associated with a causal variant. Furthermore, if the causal variant is independent of the exposure, then (4) shows that a marker-exposure interaction implies a causal variant-exposure interaction, at least under linear or log-linear regression (notably, this does not apply to logistic regression, although for rare outcomes β_{MX} will approach 0). However, if there is dependence between the causal variant and the exposure, then a marker-exposure interaction does not imply a causal variant-exposure interaction. Therefore, tests of marker-exposure interaction are not valid tests of interaction between causal variant and exposure.

We illustrate this with a numerical example. Consider a biallelic marker with population minor allele frequency (MAF) 0.1. The risk allele of the causal variant is present on half the chromosomes with the minor marker allele, but on no other chromosomes. So the MAF of the causal variant is 0.05, and the two loci are in linkage disequilibrium ($D' = 1$, $r^2 = 0.47$). If the risk allele has risk ratio 2, then assuming multiplicative risks and Hardy-Weinberg equilibrium, some simple calculations give the risk ratio for the marker as 1.5 (Table 1 and Appendix A).

Now consider a binary environmental exposure such that the risk ratio for the causal variant on the exposure is 1.5. No main effect of the exposure is assumed, although this does not matter in this example. Assuming that the quantities in Table 1 apply to unexposed subjects, some further calculations give the risk ratio for the marker as 1.6 in the exposed and 1.5 in the unexposed subjects (Table 2 and Appendix A). This reveals an interaction between the marker and the exposure on the risk of disease,

Table 1. Example Haplotype Frequencies for Disease and Marker Loci and Calculation of the Risk Ratio of the Marker

Frequency	D = 0	D = 1	Total	Pr(Y = 1,D,M)	D = 0	D = 1	Total	RR(M)
M = 0	0.9	0	0.9		0.9	0	0.9	1.0
M = 1	0.05	0.05	0.1		0.05	0.1	0.15	1.5

Abbreviations are as follows: D, allele at causal variant; M, allele at marker locus; Y, disease phenotype. Risk ratio (RR) of D = 1 is 2. Pr(Y = 1,D,M) is relative to a baseline that cancels in the marker risk ratio; see Appendix A for details of calculations.

although there is none for the causal variant. We regard this interaction as spurious, because it does not correspond to an interaction at the causal variant.

The spurious interaction arises from imperfect LD between the marker and causal variant, causing a misclassification error that differs between cases and controls, owing to the main effect of the causal variant, but that also differs between exposed and unexposed subjects, owing to the causal variant-exposure association. It is important to note that the spurious interaction cannot be removed by transformation of variables, as can be done in other cases,[4] but is a direct result of measurement error of the causal variant. It does not depend on the mechanism of gene-environment dependence, of which Figures 1, 2, and 3 show a few examples, but arises from simple algebra of the statistical model.

A particular difficulty is that the bias depends on the causal variant-exposure association, which cannot be assessed from the marker data. Indeed, the marker might not show association with the exposure, even if the causal variant is associated with both. Even if the marker is associated with the exposure, it is unclear whether or not the causal variant would be independent of the exposure after conditioning on the marker, as required by the lemma. Therefore, any test of marker-environment interaction is potentially suspect because it cannot be determined from the data whether the causal variant is associated with the exposure, conditional on the marker.

Across the thousands of markers included in GWASs and targeted array studies, it is likely that some will be in LD with causal variants that are associated with the exposure of interest. The fact that few gene-environment interactions have been reported suggests that the magnitude of the bias is small. Indeed, under a typical scenario for current GWASs in which marker and causal variant have the same MAF of 0.2, their correlation is $r^2 = 0.8$, and the causal variant has odds ratio 1.1 with both disease and a binary exposure, then similar calculations to those in Table 2 give the interaction odds ratio for the marker as 1.000433. More than a million cases and controls would be needed to detect this effect with 80% power at $p < 0.05$.[28]

However, higher interaction odds ratios can arise if the causal variant and marker have differing MAF. As an example, the marker rs10235235 observed at the *CYP3A* locus[15] has MAF 0.09 in women of European ancestry and odds ratios for breast cancer of 0.979 (95% CI: 0.915–1.047) and 0.906 (95% CI: 0.864–0.950) in women with age at menarche $\leq$12 years and >12 years, respectively. This gives an interaction odds ratio of 1.08 (95% CI: 0.990–1.176) (the original study used a finer categorization of age at menarche, leading to a significant interaction). Assume that the marker and causal variant have the maximum correlation given their MAFs (i.e., D′ = 1). Treating the odds ratios as risk ratios, we can use the approach shown in Table 2 to solve for the causal risk ratios on disease and on exposure that lead to the observed marker risk ratios, given a fixed causal MAF. For causal MAF of 0.05, the observed marker effects can arise from causal risk ratio 0.831 on disease and 0.116 on exposure. This seems unlikely because such a strong effect (0.116 = 1 / 8.62) would probably be detected by a linkage study, but this region was not identified in the largest linkage scan for age at menarche.[29] Similarly, a causal MAF of 0.01 implies a causal risk ratio of 0.155 (= 1 / 6.45) on disease and 0.208 on exposure, which again seems strong considering the lack of evidence of linkage to breast cancer in this region. However, a causal MAF of 0.02 implies causal risk ratios of 0.577 (= 1 / 1.73) on disease and 0.187 (= 1 / 5.36) on exposure, which is more plausible. Therefore, our observed marker interaction is compatible with a low-frequency causal variant with strong main effects but no interactions. Although common SNPs are generally expected to tag common causal variants,[30] the possibility of a low-frequency causal variant suggests caution in claiming a gene-environment interaction in this case.

We have focused on gene-environment interaction, but gene-gene interaction is likewise of high interest and is also susceptible to this problem. There, a spurious interaction arises if two causal variants are in LD and at least one is measured with error, such as by a marker SNP. Recently, Hemani et al.[31] have reported numerous *cis* interactions between marker SNPs on gene expression levels. However, many of the interactions can be explained by single variants in LD with both of the interacting markers (A.R. Wood, personal communication). In those cases the two causal variants are one and the same: of course a single variant is in LD, but cannot interact, with itself.

Measurement error can arise not only from a marker in LD, but also from the numerical coding of the genotype. If, for example, the true effect of a causal variant is dominant, but it is coded as additive and the linear

Table 2. Example Haplotype Frequencies and Calculation of the Marker Risk Ratio in Unexposed and Exposed Subjects

Pr(D,M|X = 0)	D = 0	D = 1	Total	Pr(Y = 1,D,M|X = 0)	D = 0	D = 1	Total	RR(M|X = 0)
Unexposed								
M = 0	0.9	0	0.9		0.9	0	0.9	0.9/0.9 = 1
M = 1	0.05	0.05	0.1		0.05	0.1	0.15	0.15/0.1 = 1.5
Exposed								
M = 0	0.9	0	0.9		0.9	0	0.9	0.9/0.9 = 1
M = 1	0.05	0.075	0.125		0.05	0.15	0.2	0.2/0.125 = 1.6

Abbreviations are as follows: D, allele at causal variant; M, allele at marker locus. Risk ratio (RR) of D = 1 is 2 on disease Y and 1.5 on exposure X. Pr(Y = 1,D,M|X) are relative to baselines that cancel in the marker risk ratio; see Appendix A for details of calculations.

model is otherwise correct, then the miscoding could also lead to a spurious interaction term. Furthermore, use of imputed rather than directly measured genotypes also creates measurement error, particularly for causal variants with strong effects because imputation is usually performed assuming no association with the outcome. Finally, exchanging the roles of gene and environment reveals that measurement error in the exposure could also create a spurious interaction even if the genotype is accurately measured.

The spurious interactions we have described are not a serious problem when the aim is to construct epidemiological models of risk, perhaps for disease prediction, in which case model fit may be improved by interaction terms. The real difficulty is with inference of biological interaction from statistical models, and

$$\frac{\Pr(Y=1\,|\,M=1,X=0)}{\Pr(Y=1\,|\,M=0,X=0)}=\frac{\sum_d R_d\Pr(d,M=1\,|\,X=0)\sum_d \Pr(d,M=0\,|\,X=0)}{\sum_d R_d\Pr(d,M=0\,|\,X=0)\sum_d \Pr(d,M=1\,|\,X=0)}.$$

our observations add to established concerns over the interpretation of statistical interactions that are model dependent.[3] We believe that additional caution is required in the interpretation of gene-environment interactions, to allow for the possibilities of gene-environment dependence and imperfect LD between marker and causal variant. We suggest that sensitivity analyses such as that described above ought to be routinely performed to reduce the possibility of false positive reports of interaction.

Appendix A

Marker Risk Ratio in Unexposed

The frequency of haplotype D,M in unexposed subjects is $\Pr(D,M|X=0)$. The joint probability of an affected subject and haplotype D,M is

$$\begin{aligned}\Pr(Y=1,D,M\,|\,X=0)&=\Pr(Y=1\,|\,D,M,X=0)\\&\quad\times\Pr(D,M\,|\,X=0)\\&=R_D\Pr(Y=1\,|\,D=0,M,X=0)\\&\quad\times\Pr(D,M\,|\,X=0)\end{aligned}$$

where R_D is the causal variant risk ratio. The risk of disease in an unexposed subject with marker M is

$$\begin{aligned}\Pr(Y=1\,|\,M,X=0)&=\Pr(Y=1,M\,|\,X=0)/\Pr(M\,|\,X=0)\\&=\sum_d\Pr(Y=1,D=d,M\,|\,X=0)\Big/\\&\quad\sum_d\Pr(D=d,M\,|\,X=0).\end{aligned}$$

The risk ratio for the marker is then

Marker Risk Ratio in Exposed

Note that

$$\begin{aligned}\Pr(Y=1,D,M\,|\,X=1)&=\Pr(Y=1,X=1\,|\,D,M)\\&\quad\times\Pr(D,M)/\Pr(X=1)\\&=S_D\Pr(Y=1,X=0\,|\,D,M)\\&\quad\times\Pr(D,M)/\Pr(X=1)\\&=S_D\Pr(Y=1,D,M\,|\,X=0)\\&\quad\times\Pr(X=0)/\Pr(X=1)\end{aligned}$$

where S_D is the risk ratio of D on the exposure X.

Also

$$\begin{aligned}\Pr(D,M\,|\,X=1)&=\Pr(X=1\,|\,D,M)\Pr(D,M)/\Pr(X=1)\\&=S_D\Pr(X=0\,|\,D,M)\Pr(D,M)/\Pr(X=1)\\&=S_D\Pr(D,M\,|\,X=0)\Pr(X=0)/\Pr(X=1)\end{aligned}$$

So the marker risk ratio in the exposed is

$$\frac{\Pr(Y=1\,|\,M=1,X=1)}{\Pr(Y=1\,|\,M=0,X=1)}=\frac{\sum_d S_d R_d \Pr(d,M=1\,|\,X=0)\sum_d S_d \Pr(d,M=0\,|\,X=0)}{\sum_d S_d R_d \Pr(d,M=0\,|\,X=0)\sum_d S_d \Pr(d,M=1\,|\,X=0)}$$

The above calculations may be performed using Table S1 available online.

Supplemental Data

Supplemental Data include one spreadsheet and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2014.07.014.

Acknowledgments

We acknowledge support from the MRC (G1000718 and K006215), Breakthough Breast Cancer, and NHS funding to the NIHR Biomedical Research Centre. We thank the Breast Cancer Association Consortium (BCAC) for access to the data published in Johnson et al.[15] BCAC is funded by Cancer Research UK (C1287/A10118, C1287/A12014) and by the European Community's Seventh Framework Programme under grant agreement number 223175 (HEALTH-F2-2009-223175) (COGS).

Received: July 2, 2014
Accepted: July 31, 2014
Published: August 21, 2014

References

1. Hunter, D.J. (2005). Gene-environment interactions in human diseases. Nat. Rev. Genet. *6*, 287–298.
2. Dempfle, A., Scherag, A., Hein, R., Beckmann, L., Chang-Claude, J., and Schäfer, H. (2008). Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur. J. Hum. Genet. *16*, 1164–1172.
3. Clayton, D.G. (2009). Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet. *5*, e1000540.
4. Thompson, W.D. (1991). Effect modification and the limits of biological inference from epidemiologic data. J. Clin. Epidemiol. *44*, 221–232.
5. Garcia-Closas, M., Rothman, N., and Lubin, J. (1999). Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. Cancer Epidemiol. Biomarkers Prev. *8*, 1043–1050.
6. Murad, H., and Freedman, L.S. (2007). Estimating and testing interactions in linear regression models when explanatory variables are subject to classical measurement error. Stat. Med. *26*, 4293–4310.
7. Hein, R., Beckmann, L., and Chang-Claude, J. (2008). Sample size requirements for indirect association studies of gene-environment interactions (G x E). Genet. Epidemiol. *32*, 235–245.
8. García-Closas, M., Thompson, W.D., and Robins, J.M. (1998). Differential misclassification and the assessment of gene-environment interactions in case-control studies. Am. J. Epidemiol. *147*, 426–433.
9. Greenwood, D.C., Gilthorpe, M.S., and Cade, J.E. (2006). The impact of imprecisely measured covariates on estimating gene-environment interactions. BMC Med. Res. Methodol. *6*, 21.
10. Piegorsch, W.W., Weinberg, C.R., and Taylor, J.A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat. Med. *13*, 153–162.
11. Mukherjee, B., and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics *64*, 685–694.
12. Gauderman, W.J., Zhang, P., Morrison, J.L., and Lewinger, J.P. (2013). Finding novel genes by testing G × E interactions in a genome-wide association study. Genet. Epidemiol. *37*, 603–613.
13. Wason, J.M., and Dudbridge, F. (2012). A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. Am. J. Hum. Genet. *90*, 760–773.
14. Dai, J.Y., Kooperberg, C., Leblanc, M., and Prentice, R.L. (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. Biometrika *99*, 929–944.
15. Johnson, N., Dudbridge, F., Orr, N., Gibson, L., Jones, M.E., Schoemaker, M.J., Folkerd, E.J., Haynes, B.P., Hopper, J.L., Southey, M.C., et al. (2014). Genetic variation at CYP3A is associated with age at menarche and breast cancer risk: a case-control study. Breast Cancer Res. *16*, R51.
16. Johnson, N., Walker, K., Gibson, L.J., Orr, N., Folkerd, E., Haynes, B., Palles, C., Coupland, B., Schoemaker, M., Jones, M., et al. (2012). CYP3A variation, premenopausal estrone levels, and breast cancer risk. J. Natl. Cancer Inst. *104*, 657–669.
17. Nickels, S., Truong, T., Hein, R., Stevens, K., Buck, K., Behrens, S., Eilber, U., Schmidt, M., Häberle, L., Vrieling, A., et al.; Genica Network; kConFab; AOCS Management Group (2013). Evidence of gene-environment interactions between common breast cancer susceptibility loci and established environmental risk factors. PLoS Genet. *9*, e1003284.
18. Thorgeirsson, T.E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K.P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature *452*, 638–642.
19. Spitz, M.R., Amos, C.I., Dong, Q., Lin, J., and Wu, X. (2008). The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. J. Natl. Cancer Inst. *100*, 1552–1556.
20. Saccone, N.L., Culverhouse, R.C., Schwantes-An, T.H., Cannon, D.S., Chen, X., Cichon, S., Giegling, I., Han, S., Han, Y., Keskitalo-Vuokko, K., et al. (2010). Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. PLoS Genet. *6*, 6.

21. VanderWeele, T.J., Asomaning, K., Tchetgen Tchetgen, E.J., Han, Y., Spitz, M.R., Shete, S., Wu, X., Gaborieau, V., Wang, Y., McLaughlin, J., et al. (2012). Genetic variants on 15q25.1, smoking, and lung cancer: an assessment of mediation and interaction. Am. J. Epidemiol. *175*, 1013–1020.
22. Liu, J.Z., Tozzi, F., Waterworth, D.M., Pillai, S.G., Muglia, P., Middleton, L., Berrettini, W., Knouff, C.W., Yuan, X., Waeber, G., et al.; Wellcome Trust Case Control Consortium (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat. Genet. *42*, 436–440.
23. Goldman, D., Oroszi, G., and Ducci, F. (2005). The genetics of addictions: uncovering the genes. Nat. Rev. Genet. *6*, 521–532.
24. Rehm, J., Baliunas, D., Borges, G.L., Graham, K., Irving, H., Kehoe, T., Parry, C.D., Patra, J., Popova, S., Poznyak, V., et al. (2010). The relation between different dimensions of alcohol consumption and burden of disease: an overview. Addiction *105*, 817–843.
25. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Mägi, R., et al.; MAGIC; Procardis Consortium (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. *42*, 937–948.
26. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
27. Taylor, A.E., Sandeep, M.N., Janipalli, C.S., Giambartolomei, C., Evans, D.M., Kranthi Kumar, M.V., Vinay, D.G., Smitha, P., Gupta, V., Aruna, M., et al. (2011). Associations of FTO and MC4R variants with obesity traits in Indians and the role of rural/urban environment as a possible effect modifier. J. Obes. *2011*, 307542.
28. Gauderman, W.J. (2002). Sample size requirements for association studies of gene-gene interaction. Am. J. Epidemiol. *155*, 478–484.
29. Anderson, C.A., Zhu, G., Falchi, M., van den Berg, S.M., Treloar, S.A., Spector, T.D., Martin, N.G., Boomsma, D.I., Visscher, P.M., and Montgomery, G.W. (2008). A genome-wide linkage scan for age at menarche in three populations of European descent. J. Clin. Endocrinol. Metab. *93*, 3965–3970.
30. Wray, N.R., Purcell, S.M., and Visscher, P.M. (2011). Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol. *9*, e1000579.
31. Hemani, G., Shakhbazov, K., Westra, H.J., Esko, T., Henders, A.K., McRae, A.F., Yang, J., Gibson, G., Martin, N.G., Metspalu, A., et al. (2014). Detection and replication of epistasis influencing transcription in humans. Nature *508*, 249–253.