GENOME-WIDE ASSOCIATION STUDIES

Gene–environment-wide association studies: emerging approaches

Duncan Thomas

Abstract | Despite the yield of recent genome-wide association (GWA) studies, the identified variants explain only a small proportion of the heritability of most complex diseases. This unexplained heritability could be partly due to gene–environment (G×E) interactions or more complex pathways involving multiple genes and exposures. This Review provides a tutorial on the available epidemiological designs and statistical analysis approaches for studying specific G×E interactions and choosing the most appropriate methods. I discuss the approaches that are being developed for studying entire pathways and available techniques for mining interactions in GWA data. I also explore methods for marrying hypothesis-driven pathway-based approaches with 'agnostic' GWA studies.

Marginal effects
The effects of a specific risk factor (gene or exposure) in the population as a whole, averaging over all other variables.

Genome-wide association study
A scan of the entire genome for association with a disease or trait using a standard panel of ~500,000 to 1 million haplotype-tagging SNPs.

Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, CHP-220, Los Angeles, California 90089-9011, USA.
e-mail: dthomas@usc.edu

doi:10.1038/nrg2764
Published online 9 March 2010

The term 'interaction' has various meanings in the epidemiologic literature, depending on the context (BOX 1). The focus of this Review is on gene–environment (G×E) interaction, here defined as a joint effect of one or more genes with one or more environmental factors that cannot be readily explained by their separate marginal effects. By convention in epidemiology, a multiplicative model is taken as the null hypothesis; that is, the relative risk of disease in individuals with both the genetic and environmental risk factors is the product of the relative risks of each separately. Therefore, any joint effect that differs from this prediction is considered to be a form of interaction. Other null hypotheses, such as an additive model for the excess risk, would yield different interpretations about interaction (BOX 1).

G×E interactions are worth studying for many reasons[1,2] (BOX 2), not least of which is the insights they could provide into biological pathways. If some of the unexplained heritability in genome-wide association studies (GWA studies) is due to interactions then — rather than discovering interactions *per se* — one goal might be to use interactions to discover novel genes that act synergistically with other factors without having demonstrable marginal effects[3]. Conversely, one might wish to discover environmental hazards that affect only a subpopulation of genetically susceptible individuals. For example, G×E interactions might allow the effects of the components of a complex mixture, such as air pollution, to be dissected[4]. Understanding the failure to replicate the findings of GWA studies is another goal, as it could provide insights into disease complexity by identifying sources of real heterogeneity[5,6]. Finally, taking account of G×E interactions in risk prediction models can have important implications for both public health and personalized medicine[7].

Traditionally, G×E interactions were investigated using candidate-gene studies. This research often begins with an established association with an environmental factor and proceeds to explore genes in pathways that are known to metabolize them. Over time, candidate-gene studies have become more elaborate investigations of entire pathways, including all of the genes, exposures and cofactors that are thought to be involved in a particular mechanism. With the advent of GWA studies, a different philosophy has gained prominence, based on 'agnostic' searches with no prior hypotheses. Understandably, most reports have focused on genetic main effects, but they are now increasingly directed at gene–gene (G×G) interactions[8]. Although many GWA studies have not collected data on environmental factors, some are based on epidemiologic cohort studies or case–control studies (TABLE 1) that have well-characterized exposure information and could be scanned for novel G×E interactions. Such scans for G×G and G×E interactions have been viewed as agnostic. Recently, however, there has been an intriguing convergence of the two philosophies: external pathway knowledge has been used to inform the analysis of GWA data to better detect signals that do not achieve genome-wide significance[9], and patterns of interaction effects have been mined from GWA data to discover novel pathways[10].

In the current post-GWA era, the focus is on integrating findings from the vast body of data that has been generated through large consortia. A key feature

© 2010 Macmillan Publishers Limited. All rights reserved

Gene–environment-wide interaction study
A scan of the entire genome for interactions with various environmental exposures.

Ecologic-level study
An observational epidemiology study that relies on comparisons of aggregate disease rates across groups in relation to aggregate exposure information rather than comparisons between individuals.

of this next phase should be a renewed focus on G×E interactions, but this will require careful consideration of epidemiologic study design, exposure assessment and methods of analysis, with particular attention to harmonization of these features across the consortia. Another key feature is the integration of GWA data with external biological knowledge from 'omics' databases.

I first discuss some of the challenges facing investigators studying environmental factors. Next, I provide a tutorial for the various types of study designs and analytical methods for studying G×E interactions in different contexts, ranging from specific interactions to more extensive biological pathways to GWA studies ('gene–environment-wide interaction studies' (GEWI studies))[11]. I discuss various ways that external data can be exploited in these types of analyses. Finally, I discuss some emerging directions and needs for making further progress.

Box 1 | Types of interaction

Statistical interaction
A departure from a pure main effects model — for example, additive or multiplicative effects for disease risk, or natural or logarithmic effects for continuous traits.

Quantitative interaction
A form of statistical interaction in which the effects of one factor go in the same direction at different levels of the other, but differ in magnitude. Lack of interaction on one scale necessarily implies interaction on other scales. For example, compared with non-carriers, carriers of rare deleterious mutations in ataxia telangiectasia mutated (*ATM*) have a more-than-multiplicative increased risk of second primary breast cancers following radiotherapy, although radiation risks are increased in both genotypes and carrier risks are increased in both exposure groups[159].

Qualitative interaction
Forms of statistical interaction in which: the effects go in opposite directions (for example, exposure is deleterious in carriers and protective in non-carriers, and vice versa); there is an increased effect only in the presence of both the environmental factor and the susceptible genotype; the effect of genotype is present at only one level of the environment; or the effect of the environment is present in only one genotype. Such interactions do not depend upon the choice of scale. For example, *in utero* tobacco smoke exposure seems to have an effect on asthma and wheeze only in children with the glutathione *S*-transferase mu 1 (*GSTM1*)-null genotype, and vice versa[160]. Opposite effects of a defensin-β1 (*DEFB1*) haplotype on asthma were seen between women and girls or between girls and boys, which suggests an interaction with some aspect of the 'internal environment'[161].

Public health synergy
A disease burden that is attributable to exposure to two or more risk factors and that is greater than the sum of the excess risks from each factor alone. For example, the population burden of gastric cancer attributable to the combination of *Helicobacter pylori* infection and interleukin-1 susceptibility alleles is greater than the sum of their separate contributions[162].

Biological interaction
An effect of one factor that depends upon the presence or absence of another[163]. For example, GST genes are inducible by oxidative stress caused by radicals and oxidants in air pollution, and myeloperoxidase levels are increased in the respiratory extrathelial lining fluid by ozone-induced inflammation[52]. This concept generally applies at the cellular or molecular level, but may have implications for statistical interactions at the whole-organism or population level.

Public health and biological interactions lead to an additive risk model as the natural null hypothesis[164], although in epidemiology the multiplicative model is more commonly used. Various authors[25,165–167] have offered classifications of different types of gene–environment interactions, including qualitative interactions (crossing, no effect of environment in those not genetically susceptible, no effect of genotype in the unexposed, and so on) and quantitative interactions.

Challenges to G×E studies

Whatever study design is used, the major challenges to the success of a G×E study — in addition to the usual challenges for genetic association studies that have been thoroughly discussed elsewhere — are exposure assessment, sample size and heterogeneity.

Exposure assessment. Many environmental factors are multidimensional; air pollution, for example, is a complex mixture of gases and particles with differing biological effects. Most environmental agents have degrees of exposure intensity that usually vary over time. Even if an exposure is not time-dependent, the resulting disease risk is likely to be modified by temporal factors, such as age at exposure or duration of exposure[12]. Seldom are accurate measurements of exposure over a lifetime available on all participants in a large epidemiologic study, but more detailed information may be obtainable on a stratified subsample to allow correction for measurement error[13]. Exposures may not even be measured on individuals, but assigned on the basis of ecologic-level studies or a prediction model. Two-phase case–control designs (BOX 3) that leverage readily available exposure surrogates to select individuals for more in-depth exposure assessment and/or genotyping might be used. Uncertainties in exposure assignments can be large and can lead to unpredictable biases, particularly if they differ with respect to disease, as well as induce spurious interactions[9]. Although methods of correction for exposure or genotype measurement errors are well established for main effects, they have seldom been applied to interaction analyses[14,15]. In general, however, interactions are less likely to be biased than main effects unless the measurement errors are differentially related to both exposure and genotype.

Sample size and power. Sample-size requirements for G×E studies can be enormous. A useful rule of thumb is that the detection of an interaction requires a sample size at least four times larger than that required for the detection of a main effect of comparable magnitude[16]. Sample sizes in the thousands of cases are typically needed for G×E analyses in candidate-gene studies, and tens of thousands are needed in GWA studies because of the more stringent significance levels required (see Supplementary information S1 (figure)). In addition to study design, the key determinants of power or sample-size requirements are the prevalence of the exposure (or its distribution if continuous), the allele frequency, the mode of inheritance, the interaction odds ratio $OR_{G\times E}$ (and to a lesser extent the odds ratios for the main effects), the significance level and the desired power. Several programs for sample size and power calculations are freely available, notably Quanto[17] and POWER[18]. It is likely that at least some of the poor track record of replicating claims of G×E interactions is due to underpowered studies in the initial discovery or replication attempts[19–21]. This has led some to suggest that the search for interactions is not worthwhile, as genes involved in interactions are more likely to be detected through their marginal effects[22]. Nevertheless, a range of interaction effect sizes can

© 2010 Macmillan Publishers Limited. All rights reserved

Box 2 | Current and potential uses of gene–environment interactions

- ***Understanding biological mechanisms and pathways.*** For example, the interaction of tobacco smoking, hair dyes and various occupational exposures with the *N*-acetyltransferase 2 (*NAT2*) gene in bladder cancer suggests a role for aryl amines[58]. Various pathway-based analyses of significant hits from genome-wide association (GWA) studies have yielded insights into underlying mechanisms of disease, but to date no analyses seem to have exploited gene–environment interactions in a gene–environment-wide interaction study.
- ***Identifying novel genes acting through interactions that are manifested by their marginal effects.*** In GWA studies in particular, these interactions could provide an explanation for some of the 'missing heritability'. GWA scans currently underway include those searching for genes that confer susceptibility to air pollution in childhood asthma or to ionizing radiation in second breast cancers, and for dietary factors that confer susceptibility to colorectal cancer.
- ***Understanding heterogeneity in results across studies caused by differences in exposure distributions.*** A meta-analysis of *NAT2* and glutathione *S*-transferase mu 1 (*GSTM1*) associations in bladder cancer[168] revealed some between-study heterogeneity in main effects, but found that the smoking × *NAT2* interaction was robust and that there was no *GSTM1* × smoking interaction.
- ***Identifying environmental factors that affect only a subgroup of genetically susceptible individuals.*** For example, maternal smoking during pregnancy seems to cause asthma only in children with the *GSTM1* null genotype[160].
- ***Dissecting the effects of complex mixtures (such as air pollution) into components that are metabolized by different genes.*** For example, the interaction between red meat consumption and *NAT2* in colorectal cancer suggests that the heterocyclic amines generated during cooking are the responsible agents[4].
- ***Establishing environmental regulation aimed at setting standards to protect the most vulnerable individuals.*** Although the US Environmental Protection Agency currently takes identifiable susceptible population subgroups (for example, children, the elderly and asthmatics) into account when setting standards, it has so far limited the use of genetic data to understanding mechanisms[169]; the use of specific genotypes in regulation raises difficult practical and ethical concerns. However, there are some voluntary employer-sponsored screening programs for human leukocyte antigen DP (*HLA-DP*) sensitivity to beryllium[170].
- ***Predicting individual risk of disease or prognosis and potential changes in risk in relation to modifiable environmental factors.*** For example, the optimal mammographic screening interval for women with a strong family history of breast cancer may differ depending on whether they carry a *BRCA1* or *BRCA2* mutation[171]. The potentially protective or deleterious effects of folate supplementation on colorectal cancer risk could depend upon genes involved in its metabolism, such as methylenetetrahydrofolate reductase (*MTHFR*)[172].
- ***Choosing the best treatment for an individual to maximize response or minimize side effects based on genetic predisposition.*** For example, a single SNP in solute carrier organic anion transporter family, member 1B1 (*SLCO1B1*) identified in a GWA study seems to dramatically affect the risk of cardiomyopathy following treatment with statins[70].

be detected in a GWA study by testing for interaction or a genetic effect in an environmental subgroup, even when the marginal effects are not detectable (Supplementary information S1 (figure)). Despite claims that interaction in the absence of main effects is a 'ubiquitous' phenomenon in nature[23,24], most examples are found at the molecular or cellular level, and there are few convincing examples in human epidemiology. Nevertheless, there are examples of genetic effects that are apparent only in groups with the relevant environmental exposure, and of environmental factors that affect only those with the susceptible genotype (BOX 1).

Interaction odds ratio
The ratio of odds ratios for the relationship of one factor (for example, a gene) with disease across the levels of another factor (for example, an environmental exposure); as such, it is a measure of departure from a multiplicative joint effect.

Heterogeneity and replication. When comparing studies that use different exposure-assessment tools, that have different distributions or characteristics of exposure (for example, different sizes or chemical constituents of particulate air pollution across regions) or that feature different confounders (for example, co-pollutants or ethnic distributions with differing genetic background risk), the potential for true heterogeneity is magnified. If explanations can be found for such heterogeneity[5], there is an opportunity for insights about the complexity of the disease, but spurious inconsistency due to methodological or data-quality differences will just add confusion.

G×E interactions with candidate genes

Any of the standard epidemiological designs for studying the main effects of genes or environmental factors — cohort designs, case–control designs or hybrid designs, such as nested case–control designs or case–cohort designs[25–27] (TABLE 1) — can also be applied to the study of G×E interactions. The issues for choosing between the designs are similar for main effects and interactions, and include the control of confounding and other biases, the temporal sequence of exposure and disease, data quality, the ability to examine multiple end points, and the efficiency of detecting rare diseases or rare risk factors (TABLE 1). For simplicity, I treat G in this section as a single functional polymorphism, but it could represent a risk-associated haplotype, several causal variants within a gene, or a risk index composed of multiple rare variants. The same analysis techniques could be applied in any case (for example, multiple logistic regression) and the design considerations would be similar. The following non-traditional designs offer particular advantages for studying interactions.

Case-only design. One of the earliest non-traditional designs was the case-only design (or 'case–case' design)[28] (TABLE 1), which can only be used for testing interactions, not main effects. This design relies on an assumption of gene–environment independence in the source population to avoid estimating this association among controls, thereby increasing power for the test of interaction. Although this assumption would be reasonable for most exogenous exposures, such as air pollution, the case-only design will yield a biased estimate of $OR_{G\times E}$ and an elevated type I error rate if the independence assumption is violated. For example, genes involved in behavioural traits, such as addiction, might be expected to produce a causal association between G and E (a G–E association) in the general population, as is sometimes seen for the environmental factor tobacco smoking[29,30]. Other G–E associations could arise indirectly, for instance between oral contraceptives and *BRCA1* through the effect of the gene on family history — a sister of an affected case might choose to take oral contraceptives to lessen her risk of ovarian cancer[31].

Broeks *et al.*[32] used a case-only design to assess the interaction between radiotherapy (RT) for the treatment of an individual's first incidence of breast cancer and mutations in four DNA damage repair genes (*BRCA1*, *BRCA2*, *CHEK2* and ataxia telangiectasia mutated (*ATM*))

© 2010 Macmillan Publishers Limited. All rights reserved

Table 1 | **Study designs for gene–environment interactions**

Design	Approach	Advantages	Disadvantages	Settings	Examples
Basic epidemiologic designs					
Cohort	Comparison of incidence of new cases across groups defined by E and G	Freedom from most biases; clear temporal sequence of cause and effect	Large cohorts and/or long follow-up needed to obtain sufficient numbers of cases; possible biased losses to follow-up; changes in exposure may require recurring observation	Common Ds or multiple end points; commonly used in biobanks	*ITGB3* × fibrinogen in platelet aggregation in Framingham cohort[154]
Case–control	Comparison of prevalence of E and G between cases and controls	Modest sample sizes needed for rare Ds; can individually match on confounders	Recall bias for E; selection bias, particularly for control group	Rare Ds with common E and G risk factors	*CYP1A2*, *NAT2*, smoking and red meat in colorectal cancer[57]
Case-only	Test of G–E association among cases, assuming G–E independence in the source population	Greater power than case–control or cohort	Bias if G–E assumption is incorrect	G×E studies in which G–E independence can be assumed	Radiotherapy × DNA repair genes in second breast cancers[32]
Randomized trial	Cohort study with random assignment of E across individuals	Experimental control of confounders	Prevention trials for D incidence can require very large sample sizes	Experimental confirmation for chronic effects	Albuteral and *B2AR* in asthmatics[126]
Crossover trial	Exposes each individual to the different Es in random order	Experimental control of confounders; within-individual comparisons	Small sample sizes; only low doses possible if E is potentially harmful	Experimental confirmation for acute effects	Immunologic marker changes following allergen and diesel exhaust particle exposure[124]
Hybrid designs					
Nested case–control	Selection of matched controls for each case from cohort members who are still D-free	The freedom from bias of a cohort design combined with the efficiency of a case–control design; simple analysis	Each case group requires a separate control series	Studies within cohorts requiring additional data collection	Antioxidants × MPO in breast cancer[155]
Case–cohort	Unmatched comparison of cases from a cohort with a random sample of the cohort	Same advantages as nested case–control; the same control group can be used for multiple case series	Complex analysis	Studies within cohorts with stored baseline biospecimens	*APOE* and smoking for CHD in Framingham offspring cohort[156]
Two-phase case–control	Stratified sampling on D, E and G for additional measurements (for example, biomarkers)	High statistical efficiency for subsample measurements	Complex analysis	Substudies for which outcome and predictor data are already available	*GST* genes and tobacco smoking in CHD[47]
Counter-matching	Matched selection of controls who are discordant for a surrogate for E	Permits individual matching; highly efficient for E main effect and G×E interactions	Complex control selection	Substudies in which a matched design is needed	Radiotherapy × DNA repair genes in second breast cancers[49]
Joint case-only and case–control	Bayesian compromise between case-only and case–control comparisons	Power advantage of case-only combined with robustness of case–control	Some bias when G–E association is moderate	G×E studies for which G–E independence is uncertain	*GSM1*, *NAT2*, smoking and diet in colorectal cancer[34]
Family-based designs					
Case–sibling (or –cousin)	Case–control comparison of E and G using unaffected relatives as controls	More powerful than case–control for G×E; immune to population stratification bias	Discordant sibships difficult to enroll; overmatching for G main effects	Populations with potential substructure	*GSTM1* × air pollution in childhood asthma[17]
Case–parent triad	Comparison of Gs for cases with Gs that could have been inherited from parents, stratified by case's E	More powerful than case–control for G×E; immune to population stratification bias for G main effects	Difficult to enroll complete triads; possible bias in G×E if G and E are associated within parental mating types	Substructured populations, particularly for Ds of childhood	*TGFA* × maternal smoking, alcohol and vitamins in cleft palate[157]
Twin studies	Comparison of D concordance between MZ and DZ pairs in different environments	No genetic data required; can be extended to include half-siblings, twins reared together or apart, or to compare discordant pairs on measured G and E	Used mainly to identify interactions with unmeasured genes; assumption of similar E between MZ and DZ pairs	Exploratory studies of potential for G×E before specific genes have been identified	Concordance of insulin levels in relation to non-genetic variation in obesity[158]

© 2010 Macmillan Publishers Limited. All rights reserved

Table 1 (cont.) | **Study designs for gene–environment interactions**

Design	Approach	Advantages	Disadvantages	Settings	Examples
GWA designs					
Two-stage genotyping	Use of high-density panel on part of a case–control sample to select a subset of SNPs with suggestive Gs or G×E interaction for testing; the SNPs are tested using a custom panel in an independent sample, with joint analysis of both samples	Highly cost efficient	Only part of sample has GWA genotypes	GWA studies for which complete SNP data on all subjects is not needed	None identified
Two-step interaction analysis	Preliminary filtering of a GWA scan for G–E association in combined case–control sample, followed by G×E testing of a selected subset	Much more powerful for G×E or G×G interactions than a single-step analysis	Can miss some interactions	GWA studies with complete SNP data and focus on G×E	G × *in utero* tobacco in childhood asthma
DNA pooling	Comparison of allelic density in pools of cases and controls stratified by E, followed by individual genotyping	Highly cost efficient	Technical difficulties in forming pools and assaying allelic density; limited possibilities for testing interactions	GWA studies for which an initial scan is severely limited by cost	None identified

APOE, apolipoprotein E; *B2AR*, adrenergic β2 receptor (also known as *ADRB2*); CHD, coronary heart disease; *CYP1A2*, cytochrome P450 family 1, subfamily A, polypeptide 2; D, disease; DZ, dizygotic; E, environment; G, gene; G×E interaction, gene–environment interaction; G–E association, causal association between gene and environment; *GST*, glutathione *S*-transferase; *GSTM1*, glutathione *S*-transferase mu 1; GWA, genome-wide association; *ITGB3*, integrin-β3; MPO, myeloperoxidase; MZ, monozygotic; *NAT2*, *N*-acetyltransferase 2; *TGFA*, transforming growth factor-α.

on the subsequent risk of contralateral breast cancer (CBC). Among RT+ cases, there was a 2.2-fold higher prevalence of germline mutations in one or more of these genes than among RT– cases. Here it seems unlikely that genotypes would have affected the choice of treatment, except perhaps indirectly through tumour characteristics or stage at diagnosis (factors that could be adjusted for).

It is tempting to begin by testing for G–E association in controls and then decide whether to use the case-only test (for greater power if there is no G–E association) or the case–control test (for greater validity if there is). However, this naive procedure leads to biased tests and estimates because it fails to take proper account of this two-step inference procedure[33]. More appropriate empirical Bayes[34] or Bayes model averaging[35] approaches have been developed that essentially provide weighted averages of the case-only and case–control estimators, yielding an acceptable trade-off between bias and efficiency. For example, Mukherjee *et al.*[34] re-analysed data on glutathione *S*-transferase mu 1 (*GSTM1*) and *N*-acetyltransferase 2 (*NAT2*) genotypes in relation to smoking and dietary factors. They found a strong association between *NAT2* and smoking, so their empirical Bayes estimate of the interaction between the two was closer to the case–control estimate than to the case-only one, which was in the opposite direction. However, there was no association between *GSTM1* and fruit consumption, so the empirical Bayes estimate of that interaction was similar to both the case–control and case-only estimates, but took advantage of the smaller standard error of the latter.

Family-based association tests. Family-based association tests — case–parent triad designs[36], case–sibling designs[37], designs using extended pedigrees[38], and modified segregation analyses[39] (TABLE 1) — are appealing because they avoid bias from population stratification, but are generally less powerful for testing main effects than case–control studies using unrelated controls. However, they can be more powerful for testing G×E interactions if relatives' exposures are not too highly correlated[37]. Population stratification can bias G×E interactions only if the substructure is related to the gene and the environmental factor differentially — that is, there are different ancestry–genotype associations in exposed and unexposed individuals — which seems unlikely. The case–parent triad design requires exposure information only on the cases (although it does require surviving parents for genotyping, making it more suitable for early-onset diseases) and entails a comparison of genetic relative risks between exposed and unexposed cases. The discordant sibship design requires exposure information on all cases and controls and uses standard conditional logistic regression tests of interaction. Twin studies[40] (TABLE 1) and joint segregation and linkage analysis[41–44] can also be used for testing the existence of G×E interactions with unknown genes or specific regions[25].

Two-phase case–control design. Two other novel designs use different ways of selecting controls to improve the power for detecting either main effects or interactions. The two-phase case–control design[45] is useful when a surrogate for exposure is readily available but additional expensive data collection is required to retrieve data on exact doses, confounders or modifiers[46]. (Note that the kinds of two-phase sampling designs described here are fundamentally different from the two-stage genotyping designs for GWA studies described below and in BOX 3.) These designs entail independent subsampling on the basis of disease status and of the exposure surrogate variable from a first-phase case–control or cohort study. Data from both phases are combined in the analysis, with appropriate allowance for the biased sampling in phase two. The optimal design entails over-representing the rarer cells, typically the exposed cases. Although most applications have focused on the use of the two-phase case–control design for improving exposure

Confounder
A spurious association between a risk factor (a gene, exposure or interaction) and disease induced by the joint associations of some other variable with the risk factor and the disease that are independent of the risk factor. Confounding can also distort the magnitude of the association of a true risk factor with disease or mask it.

Gene–environment independence
The independent distribution of genotype and environment in the source population.

Empirical Bayes
A technique for estimating the effects of each component of a large ensemble of related variables by assuming the ensemble has some common distribution and estimating the parameters of that distribution. Empirical Bayes estimators typically have better prediction error than estimating each one separately.

Bayes model averaging
A technique for accounting for uncertainty about the correct model form (for example, the selection of variables to include in a multiple regression model) by averaging the effects of each possible variable over the set of all plausible models.

© 2010 Macmillan Publishers Limited. All rights reserved

Box 3 | Designs for genome-wide interaction scans

Although any of the designs for studying gene–environment (G×E) interactions with single genes could be used for genome-wide association (GWA) studies that include interactions (gene–environment-wide interaction studies), the following five have the potential to greatly improve power or cost-efficiency.

Two-phase case–control designs

These combine GWA SNP data, stratified jointly by disease and exposure, from a subsample of a large epidemiologic case–control or cohort study with the data on exposure (and possibly established genes) from the parent study, and adjustments are made to account for the biased sampling. For example, Li *et al.*[47] compared coronary heart disease cases with a stratified subcohort based on age, gender and carotid intima thickness and found an interaction between smoking and the glutathione *S*-transferase theta 1 (*GSTT1*)-null genotype.

Two-stage genotyping designs

These designs use high-density genotyping chip or array technology to assay hundreds of thousands or over a million SNPs from a random sample of cases and controls. The most promising SNPs are then selected, based on their main effects and interactions, for custom genotyping in the remainder of the sample. The final analysis combines the information on the selected SNPs and environmental factors from both samples.

Two-step analyses

In two-step analyses the multiple comparisons penalty for looking at all possible interactions within a sample with complete GWA SNP data is reduced by restricting the final analysis to only a subset of the possible interactions based on a preliminary filtering step. Two approaches to this filtering have been suggested. The first approach involves restricting comparisons to the subset of gene and environment variables that show marginal effects at a liberal significance level[95]. The second approach involves testing all possible causal associations between G and E (G–E associations) in the combined case–control sample and then testing only those combinations for G×E interaction, using a standard case–control comparison[99] (FIG. 1).

Joint case-only and case–control designs

In these designs the empirical Bayes method or Bayes model averaging is applied to all possible interactions in combined case-only and case–control tests.

DNA pooling

Here, pools of DNA from cases and controls, stratified by exposure, are tested for differences in allele frequency, followed by individual genotyping in the same or new samples.

characterization for main effects or for better control of confounding, it can also be highly efficient for studying interaction effects. For example, Li *et al.*[47] used a two-phase design nested within the Atherosclerosis Risk in Communities (ARIC) study to study the interaction among *GSTM1* or glutathione *S*-transferase theta 1 (*GSTT1*), cigarette smoking and the risk of coronary heart disease. Their sampling scheme was not fully efficient for addressing this particular question because it stratified only on intima media thickness, not smoking, and only for the controls, and it did not exploit the information from the original cohort in the analysis. Re-analyses of other data from the ARIC study[48] showed the considerable improvement in efficiency that can be obtained by using the full cohort information.

Counter-matching. Counter-matching (TABLE 1) is essentially a matched variant of the two-phase design. Here, one or more controls are selected for each case on the basis of exposure so that each matched set contains the same number of exposed individuals. Another study of CBC in relation to RT and DNA damage repair genes[49] counter-matched each CBC case to two controls with unilateral breast cancer, such that each matched set contained two RT+ subjects. Radiation doses to each quadrant of the contralateral breast were then estimated and DNA was obtained for genotyping candidate DNA repair genes and for a GWA scan. Langholz[50] has shown the considerable gains in power that can be obtained, both for main effects and for interactions. In particular, for G×E interactions Andrieu *et al.*[51] showed that a 1:1:1:1 design counter-matched on surrogates for both exposure and genotype was more powerful than conventional 1:3 nested case–control designs, or 1:3 or 2:2 designs counter-matched on just one of these factors.

Approaches for candidate pathway analyses

So far I have considered interactions between one gene and one environmental factor, but most candidate gene studies are based on a conceptual model for one or more candidate pathways. For example, most of the genetic studies being done for susceptibility to the effects of air pollution on children's asthma and lung growth in the Southern California Children's Health Study have been motivated by a theoretical framework involving oxidative stress, inflammation and modifiers, such as antioxidant intake[52]. Typically, such hypotheses lead to the selection of a set of candidate genes to be studied together. How then can these data be analysed in combination to learn about the overall effect of the postulated pathway(s)?

Multifactor dimension reduction. Many exploratory methods have been developed for multivariate analysis of high-dimensional data, ranging from standard multiple regression techniques to various machine learning or pattern recognition methods[8,53,54]. Perhaps the most popular of these methods for studying interactions is multifactor dimension reduction (MDR)[8,55,56], which I applied in BOX 4 to data on a reported four-way interaction among two exposures (smoking and red meat) and two genes (cytochrome P450 family 1, subfamily A, polypeptide 2 (*CYP1A2*) and *NAT2*) in colorectal cancer[57]. Although this study is widely quoted as one of the few examples of a higher-order interaction, this analysis makes clear that the four-way interaction is not internally reproducible by cross-validation. In this instance, MDR is more useful for putting a high-dimensional interaction into context than for discovering one, and emphasizes that if two-way interactions require large sample sizes, higher-order interactions require even larger sample sizes. Nevertheless, the interaction is biologically plausible (similar replicated interactions among *NAT2*, *GSTM1*, tobacco smoking and occupational exposures have been reported for bladder cancer[58]) and is worth studying further using techniques that leverage known pathways.

Gene-set-enrichment analysis and hierarchical models. As candidate pathway studies are hypothesis-driven, it seems appropriate to carry this reasoning through to the analysis[59,60]. Two approaches that attempt to leverage external information about biological pathways are summarized below and in BOX 5. These methods, though promising, have not been widely applied to candidate-gene studies so far.

Modified segregation analysis
This analysis applies likelihood-based methods to data from a pedigree in which one or more members have genotypes available at a major gene. It derives the genotypes of untyped individuals by summing their conditional genotype probabilities using the genotypes available.

Population stratification
The phenomenon of an apparently homogeneous population that is actually composed of subgroups of individuals with distinct ancestral origins and differing allele frequencies at many loci. This leads to bias in the assessment of the significance of associations of a trait with particular loci.

© 2010 Macmillan Publishers Limited. All rights reserved

Box 4 | Multifactor dimension reduction

The table shows my reanalysis — using the multifactor dimension reduction (MDR) technique — of grouped data from Le Marchand *et al.*[44] on colorectal cancer in relation to two exposures, smoking and red meat, and the phenotypic markers of two genes, cytochrome P450 family 1, subfamily A, polypeptide 2 (*CYP1A2*) and *N*-acetyltransferase 2 (*NAT2*).

Data set	CYP1A2 activity	NAT2 acetylation	Cases/controls			
			Non-smoker		Smoker	
			Rare or medium meat	Well-done meat	Rare or medium meat	Well-done meat
Training subset (nine-tenths of the samples)	≤ median	Slow/intermediate	31/51*	15/11‡	39/44*	12/19*
		Rapid	15/23*	9/14*	25/30‡	10/12‡
	> median	Slow/intermediate	32/46*	16/19‡	16/23*	8/6‡
		Rapid	51/58‡	20/32*	9/21*	10/2‡
Testing subset (one-tenth of the samples)	≤ median	Slow/intermediate	1/6*	3/1‡	1/11*	1/3*
		Rapid	1/3*	0/1*	2/5*	0/0
	> median	Slow/intermediate	0/7*	1/0‡	0/5*	1/0‡
		Rapid	10/12‡	5/1‡	2/0‡	2/0‡

*Low-risk category. ‡High-risk category.

The proportion correctly classified in the testing subset by the rule derived from the training data for this realization is 58/85 (68.2%). Across 10 random training/testing subsets, however, the mean classification accuracy is only 49.7% (range 31.9–74.1%); this is no better than chance, due to the small numbers of subjects (12 cases, 2 controls) in the high-risk category. All possible models (combinations of genes and environmental factors) were explored using MDR, and only the main effect of smoking on colorectal cancer risk was found to be replicable.

Gene-set-enrichment analysis (GSEA)[61] (BOX 5) tests whether disease-associated genes are significantly enriched for particular pathways. Although GSEA is widely used in the analysis of gene-expression data, methods for applying it in association studies have only recently been developed[62–64] and have not yet been used for G×E studies.

Hierarchical models (BOX 5) extend traditional multiple regression methods for exploring main effects and interactions in an epidemiological data set by regressing the first-level coefficients on external data[65–67]. External information can include simple pathway indicator variables[68], genomic annotation or pathway ontologies[69], functional assays[70], *in silico* predictions of function or evolutionary conservation[71], or simulation of pathway kinetics[72,73].

The GSEA and hierarchical modelling approaches can be thought of as 'empirical' because they use external information only to guide the selection of terms to include in a model or to stabilize their estimation. These approaches do not fit strong mechanistic models directly — our understanding of the basic biology is too primitive — although there have been notable successes. Some of the earliest were stochastic models for multistage carcinogenesis[74,75], but they have not been applied to pathways involving specific genes. Other areas that have seen extensive mathematical modelling include the pharmacokinetics and pharmacodynamics of drug metabolism[76], of exposure to toxic substances[77,78] and of normal metabolism[79,80]. Although inter-individual variation in metabolic rate parameters has long been recognized, their genetic basis has only recently been incorporated into this kind of modelling[81,82].

Use of biomarkers. Even when supplemented with external information, the informativeness of epidemiological studies of chronic disease end points for the purpose of pathway analysis is limited by the dichotomous nature of the phenotype. The information content may be improved by obtaining biomarker data on some of the intermediate steps in the process. Ideally, biomarker specimens would be sampled longitudinally and before disease onset. This may be prohibitively expensive, so the two-phase case–control design samples individuals from a cohort or case–control study based on disease, exposure and genotype information[83]. Nested case–control studies in biobanks overcome the problem of reverse causation by using stored specimens and exposure information obtained at enrolment. Mendelian randomization[84,85] provides another way to avoid reverse causation by using genes (which are not subject to this problem) as instrumental variables[86] for the biomarker–disease relationship. In a randomized trial of oestrogen plus progestin, Dai *et al.*[87] used a two-phase design to assess interactions of treatment with thrombosis biomarkers. They found that interaction-effect estimates made by using their two-phase design were considerably more precise than estimates made by using the case–control study alone or by using standard two-phase estimators that do not assume G–E independence.

Mining GWA data for G×E interactions

Although the approaches described above could be used in a genome-wide context, the enormous cost, computational burden, multiple comparisons penalty and general absence of prior knowledge about most SNPs pose additional complexities. For the main effects of genes,

Joint segregation and linkage analysis
The use of family studies to estimate the parameters of a penetrance model. The parameters could include interactions between the unobserved major gene, which is linked to a marker, and environmental factors.

Multiple regression
A standard statistical technique for relating a single outcome variable to multiple explanatory variables, either all at once or using some variable selection method, such as stepwise forward selection or backward elimination.

Machine learning
Any of many data analysis techniques for mining large data sets derived from the computer science field. The techniques are not specifically based on mathematical statistics theory.

Pattern recognition
Any technique from exploratory data analysis or machine learning for discovering non-random patterns in large data sets.

First-level coefficients
In a hierarchical model, the regression coefficients (for example, log relative risks for each variable) for the subject-level data on the association between risk factors and disease. Unlike a non-hierarchical model, these coefficients are treated as random variables with distributions described in the higher level(s) of the model rather than as model parameters to be estimated directly.

Pathway indicator variables
Various types of information that can be used as predictor variables in the higher levels of a hierarchical model, specifically binary variables that indicate whether a particular gene or interaction has a role in a particular pathway.

© 2010 Macmillan Publishers Limited. All rights reserved

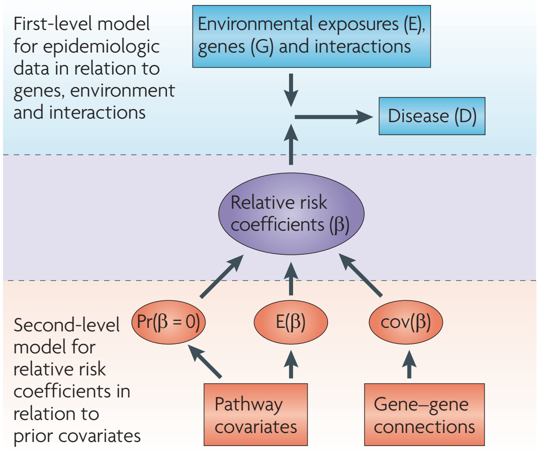
Box 5 | Pathway-based approaches for genome-wide association study analysis

Gene-set-enrichment analysis

This approach shifts the emphasis from the effects of individual SNPs to sets of genes known *a priori* to have related functions. First, each SNP is assigned to one or more genes, typically based on proximity, and a summary statistic for each gene is obtained (for example, the minimum *p* value for all SNPs assigned to it). Then genes are assigned to gene sets and the distribution of gene-specific summary statistics for each set is compared with its null distribution, typically using the Kolmogoroff–Smirnoff test. Permutation may be used to allow for the non-uniformity of the null distributions. This method seems to have been applied only to purely genetic analyses, but could be extended to the genes involved in gene–environment interactions.

Hierarchical models

This approach supplements a traditional epidemiologic analysis (for example, multiple logistic regression) with a second level in which the first-level regression coefficients are modelled in relation to a set of ‘prior covariates’ or information about connections between genes derived from external information, such as pathway or genomic databases (see the figure). This shifts the main focus of inference from the effects of specific exposures, genes or interactions to the effects of the pathways or other external predictors. It also provides more stable estimates of the individual risk factor effects by ‘borrowing strength’ from related risk factors. The first-level associations may comprise a mixture of null and non-null associations, with probability depending upon prior covariates. The prior means of the non-null effects are regressed on prior covariates, and their covariances can depend on a matrix of gene–gene connections. Rebbeck *et al.*[18] provide a discussion of various sources of prior covariate information. cov(x,y), covariance between x and y; E(x), expectation of x; Pr(x), probability of x.


First-level model for epidemiologic data in relation to genes, environment and interactions
Environmental exposures (E), genes (G) and interactions
Disease (D)
Relative risk coefficients (β)
Second-level model for relative risk coefficients in relation to prior covariates
Pr(β = 0)
E(β)
cov(β)
Pathway covariates
Gene–gene connections

various design and analysis issues have been widely discussed[88,89], so the remainder of this Review focuses on the use of GWA data for analysing G×E interactions. Two-stage genotyping designs and two-step analyses of a single-stage design (discussed below) could be applied to interaction studies (BOX 3). In contrast to the pathway-based approaches in the previous section, these novel techniques are currently applicable to GWA data.

Two-stage genotyping design. The two-stage genotyping design[90] has been extended to the GWA scale[91–94] and used to discover main effects in many studies. The design is also attractive for GEWI studies, but requires choices about how to select the SNPs to be carried forward to the second stage based on promising main effects and interactions. Any SNP for which the main effect or any of the G×E or G×G interaction tests attained the appropriately Bonferroni-corrected significance level would be chosen for inclusion in stage-two genotyping. To maximize the yield of true positives, knowledge of the distribution of the true effect sizes for each type would be required to ensure optimal selection of hits; however, reasonable bets on which hits to pursue can be made based on previous literature and calculation of the power to detect similar effects.

Two-step analysis approaches. A conventional two-step analysis of G×G interactions in a single-stage GWA study restricts the search for interactions to gene pairs for which one or both members show a marginal association. It can be more powerful than an exhaustive scan for all possible pair-wise interactions but risks missing those with no or weak marginal effects[8,95–97]. In addition, scanning for higher-order (G×G×G…) interactions is computationally unfeasible without filtering based on main effects and/or lower-order interactions. Although this filtering approach could also be applied to G×E interactions, it does not exploit the ability of the following two-step approaches to use different designs.

The case-only design is appealing for a GEWI study because it has greater power than the case–control design and because most GWA SNPs are unlikely to be correlated with environmental factors in the source population. Nevertheless, some false positives due to G–E association may occur, and even if only a small proportion of all SNPs was associated, this could represent a high proportion of all reported G×E interactions. Because any scan for interactions is likely to have been accompanied by a main effects scan, controls are probably available anyway, so it would be wasteful not to use them. (The exception would be if public controls with no environmental data, or non-comparable data, were used for the main effects scan, combining case-only information on G×E interactions with case–control information on genetic main effects[98].) Two basic approaches have been suggested for taking advantage of controls to protect against false positives while exploiting the power advantage of the case–control design. Murcray *et al.*[99] introduced a two-step analysis of a single-stage GWA study (FIG. 1) in which G–E association is first tested in the combined case and control sample and then only the most significant SNPs are tested for G×E interaction using the standard case–control test. The second general approach is the use of empirical Bayes[34] or Bayes model

Ontology
A formal system for organizing knowledge, here used in the context of biological pathways as a means of synthesizing information about the function of genes and exposures and their joint roles in disease causation.

Reverse causation
A bias in the estimation of the causal effect of a biomarker on disease when biospecimens are obtained after diagnosis. The bias occurs because the disease or its treatment alters the underlying intermediate variable or the measurement of it.

Mendelian randomization
A technique for studying the relationship between a biomarker and disease indirectly by studying the relationship of each to a gene that influences the biomarker.

Instrumental variable
In statistics, a variable that can be used to predict the value of an explanatory variable that is measured with error. The instrumental variable thereby indirectly yields an unbiased estimate of the relationship of the explanatory variable with an outcome variable.

Multiple comparisons penalty
The higher degree of statistical significance that is required for a particular association to be considered noteworthy when many possible associations are analysed simultaneously. Several adjustment methods can take account of this penalty, the best known of which is the Bonferroni correction.

Bonferroni correction
A multiple comparisons adjustment for testing at a conventional significance level. It is based on multiplying the *p* value for a specific test by the total number of tests performed, and approximately controls the overall type I error rate (the probability of at least one false positive association) at the chosen significance level if the predictors are independent.

© 2010 Macmillan Publishers Limited. All rights reserved

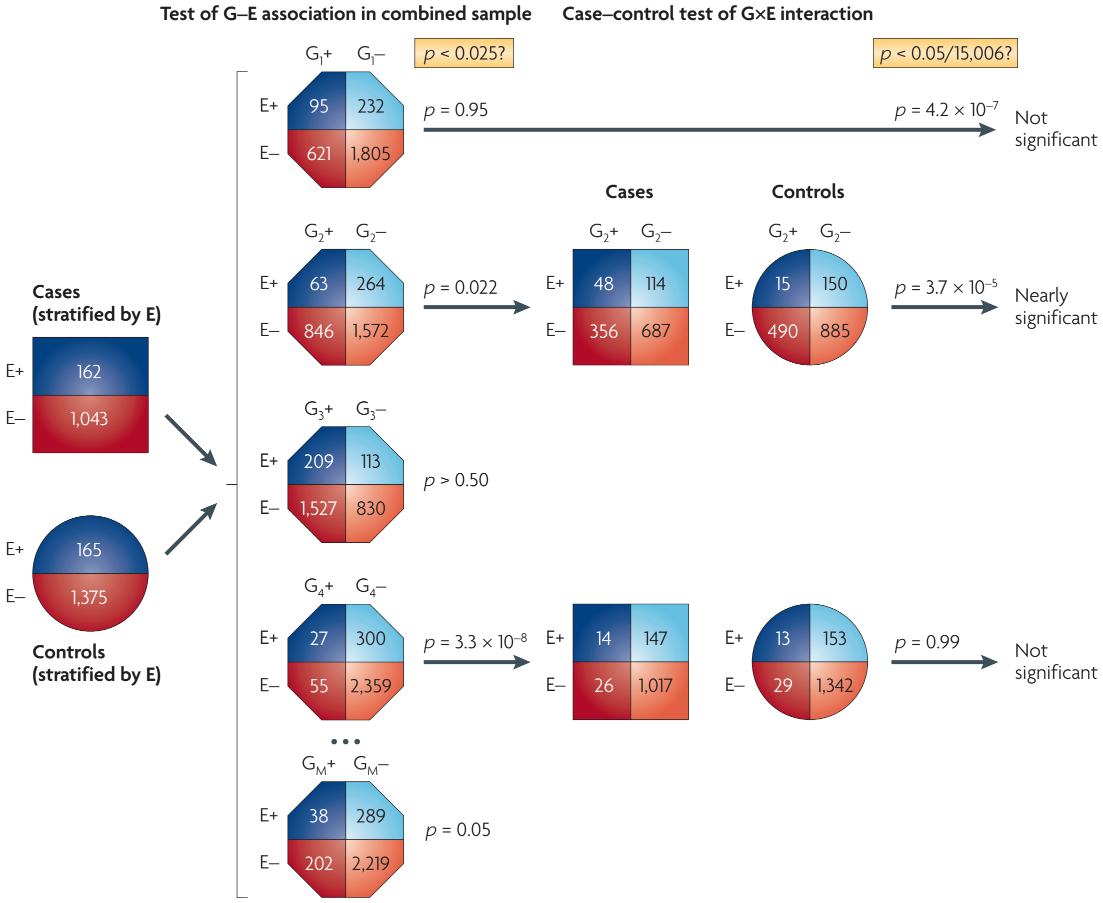


Figure 1 | **Schematic representation of the two-step gene–environment-wide interaction test.** Schematic representation of the two-step gene–environment-wide interaction (GEWI) test for gene–environment (G×E) interaction used by Murcray *et al.* (data from REF. 99). G$_1$, G$_2$, G$_3$, and so on to G$_M$ denote the genotypes at each SNP in a genome-wide association (GWA) study and E denotes a binary exposure variable. Association between gene and environment (G–E association) is tested in the combined case and control sample, and only the most significant SNPs are then tested for G×E interaction using the standard case–control test (in this example, the second and fourth rows are taken forward to the second step). Despite the dilution of the induced G–E association in the first step by the inclusion of the controls, this approach yields a second-step test that is independent of the first and therefore only needs to be corrected for the number of SNPs that are actually taken forward to the second step. They showed that the resulting procedure has dramatically better power than a conventional single-step case–control comparison. The optimal design depends only weakly on the true model parameters. For rare diseases with a 1:1 ratio, any first-stage significance level of α_1 ~ 0.0001 yields roughly similar power, although a common disease would require a much larger α_1. When this test was applied to data from the Southern California Children's Health Study for asthma, 15,006 SNPs that attained an optimized first-step threshold of $\alpha_1 = 0.025$ were identified in the first-stage test of association between SNPs and *in utero* tobacco smoke exposure in the combined case–control sample. When the second-stage case–control test was carried out on these SNPs, one nearly significant interaction (the second example in the figure) was found that would not have achieved genome-wide significance in a traditional one-step test, or been deemed significant by its main effect. This SNP shows no effect in the absence of *in utero* tobacco exposure and exposure shows no effect in non-carriers of the minor allele. The first row shows the most significant SNP×E interaction in a conventional single-stage test; in the two-step procedure, this SNP fails the first step and hence is declared not significant. The fourth row shows the most significant SNP–E association in the first step, which shows no sign of SNP×E interaction in the second step. (The marginal totals differ slightly from row to row because of missing genotypes.)

averaging[35] methods that combine the case-only and case–control estimators to provide a reasonable trade-off between validity and efficiency. Simulation studies show that these approaches can have better power than the two-step analysis over a range of modest interaction-relative risks, whereas the two-step approach is more powerful for larger interaction-relative risks.

DNA pooling. Another possible approach for saving on genotyping costs is DNA pooling (BOX 3), at least for an initial screen, to be followed by individual genotyping of promising loci[100]. Beyond the technical challenges of forming comparable pools and assaying allelic concentrations, this approach would be feasible for studies of G×E interactions only if the pools were stratified on

© 2010 Macmillan Publishers Limited. All rights reserved

the basis of exposure, therefore limiting the number of possible environmental factors that could be considered. Recent advances in DNA bar-coding[101], however, would permit the reconstruction of individual genotypes from within pools[102], thereby allowing a broader range of interaction analyses.

Prioritization of hits to pursue. One must sift through a massive number of potential 'hits' to decide which should be considered in independent replication studies, functional assays or subsequent stages of a multistage genotyping design. This decision is usually based on statistical significance, but also entails expert judgment based on the internal consistency of the results and the coherence with other knowledge (for example, the existence of other GWA associations for the same or related traits or biological pathways). Coherence has tended to be a more informal judgment, but various methods have emerged for formalizing this process. The following techniques can be viewed as well established and available for application now, although because of their novelty, there are few applications so far. See REF. 103 for an excellent review of the available techniques in the context of genetic main effects.

One of the first prioritization techniques was a weighted false discovery rate (FDR) approach[104]. This approach uses external information to prioritize some SNPs or regions while maintaining a fixed overall FDR. Bayesian versions of the FDR have also been described[105,106], as well as the use of Bayes factors[107] and empirical Bayes shrinkage[108]. GSEA and hierarchical modelling approaches are also amenable to incorporating external knowledge. Several authors[109–111] have described applications of the hierarchical Bayes modelling approach for GWA data using prior covariates extracted from genomic or pathway ontologies. Although these have focused on main effects, the methods are also applicable to GEWI studies[11], the limiting factor presently being the lack of suitable ontologies for interaction effects. Meanwhile, various ways of using GSEA or other methods of integrating pathway knowledge into GWA analyses are being discussed[9,62–64,112–116]. Few studies have explicitly included G×E interactions in formal pathway-based analyses of GWA data[117]. A promising approach entails incorporating metabolomics, as in the first GWA study of a large panel of metabolite phenotypes[118]. The authors identified associations between four enzyme-encoding genes and ratios of metabolite concentrations. These metabolic profiles were consistent with the pathways in which these enzymes are known to act.

Methods for discovering novel pathways. An emerging idea is to use Bayesian network analysis[119–121] or similar techniques to discover novel pathways. Bayesian networks have been widely used in the analysis of gene co-expression data to discover cliques of interacting loci. The starting point is usually a matrix of gene–gene correlations across multiple experimental conditions (for example, time series of synchronized cell cultures or different environmental stressors), which can be used to derive a parsimonious graphical representation of the important interactions. Unlike co-expression data, GWA data provide only a single estimate of the association between genotype and phenotype, but no information about gene–gene connections. G×G interaction analyses do, however, yield information about pairs of genes that could be mined in a similar way, as could G×E interactions. Sebastiani *et al.*[10] applied the technique to modelling the posterior probability of genotypes and exposures according to disease status, yielding graphical models that can be interpreted in terms of interactions. However, these probabilities depend on both the risk of disease given G and E (and their interactions) and the correlations among these factors, so they do not represent a pure interactome model[122]. Alternatively, a known network can be used as a prior covariance matrix for main effects or to provide prior covariates for interactions in a hierarchical model (BOX 5). Although potentially exciting, such methods have yet to be applied on a GWA scale.

Experimental validation of G×E interactions

Experimental studies offer unique promise for validating G×E interactions, as both exposure and genotypes can be carefully controlled through randomization. Model organisms are commonly used for evaluating genetic modifiers of drug response; for example, Koch and Britton[123] used selective breeding of rats on aerobic capacity to study gene–diet interactions in combination with body weight and various metabolic markers. In human challenge studies, a randomized crossover design is typically used, in which volunteers are exposed to one or more environmental exposures in random order. In one intra-nasal challenge study of allergen alone or with diesel exhaust particles, various immunological responses were measured[124]. Stratified analyses revealed that those with the *GSTM1*-null or glutathione *S*-transferase pi 1 (*GSTP1*) I/I genotypes had significantly larger increases in immunoglobulin E and histamine levels after diesel challenge. Subjects were not pre-selected on the basis of genotype, so results were limited by the relatively small numbers of subjects with the susceptible genotypes. Challenge studies nested within epidemiologic cohorts for which genotypes (and possibly various outcomes) are already available could be more powerful.

Clinical trials also allow controlled comparisons for G×E interactions and more powerful designs using two-phase sampling on various combinations of genotype, treatment, outcomes and possibly other factors[93,125]. For example, Israel *et al.*[126] performed a clinical trial of albuteral in asthmatics, matching pairs on forced expiratory volume and adrenergic β2 receptor (*B2AR*, also known as *ADRB2*) genotypes, and found a highly significant gene × treatment interaction. A case-only design nested within a clinical trial is particularly appealing for evaluating gene–treatment interactions on survival or other treatment responses, as treatment assignment is independent of genotype by virtue of randomization[127,128].

Needs for further progress

Better ontologies. The biggest barrier to integrating biological knowledge with agnostic GEWI studies data may be the lack of ontologies designed to bring together information from SNPs, genes and pathways, but also

DNA bar-coding
The addition of a unique molecular tag to each fragment of an individual's DNA so that after pooling with other DNA samples, the genotype of each individual in the pool can be reconstructed.

Coherence
The extent to which the data at hand is concordant with other types of biological knowledge, thereby reinforcing a causal interpretation.

False discovery rate
This controls the proportion of all reported positive associations that are expected to be false positives, and can be used to judge which of many associations are noteworthy.

Bayesian network analysis
A technique for developing a minimal graphical representation of the connections among a large set of variables by examining the conditional independence relationships among pairs of variables given the other variables connected to them within the graph. This technique has been widely used for the analysis of gene co-expression data.

Challenge studies
Various experimental designs for assessing the effects of a noxious agent by exposing individuals to trace amounts in a controlled setting (as in a randomized or crossover trial). For gene–environment interaction studies, the effects can be compared across subgroups with different genotypes, and the efficiency can be improved by stratified sampling based on genotype.

© 2010 Macmillan Publishers Limited. All rights reserved

their relevant environmental substrates, known relationships to disease, metabolic parameters and toxicological information. The creation of such a database is arguably one of the most important contributions of the Human Genome Epidemiology Network (HuGENet) project[129], but is highly labour-intensive because expert curation of the literature is needed. HuGENet's valuable series of reviews on specific topics[130,131] does not replace the need for a searchable database that could provide prior covariate information in a systematic and unbiased manner. Automatic literature-mining approaches[132,133] have been developed that can help to assign sets of genes to shared pathways or interaction networks. However, they are still vulnerable to bias in what is investigated and published; the current literature on G×E interactions is very sparse, highly subject to publication bias, poorly replicated and tends to reflect a 'looking under the lamp post' mentality in terms of what gets studied. Other genomic or pathway ontologies[134–136] tend to be limited to purely genetic information and are only partially useful for G×E modelling.

Environmental pathways mediated through epigenetics and other mechanisms. One of the aims of pathway-based modelling is to understand how genetic and environmental effects are mediated through intermediate events, such as changes in gene expression, epigenetic processes (such as DNA methylation)[137], somatic mutations[138] and interference by small RNAs[139]. These phenomena have been studied in relation to disease and to a lesser extent exposure[140,141], but the full pathways from genes and exposures through epigenetics to disease remain to be studied[137]. For example, the seminal observation[142] that monozygotic twins start life with identical methylation patterns but subsequently diverge suggests the effect of environmental factors and may provide a mechanism for their subsequent discordance in disease. Latent variable models could be used to treat biomarker measurements as surrogate observations of a long-term unobserved process leading to disease. Various omics technologies could provide high-dimensional measurements of intermediate processes on targeted subsamples of epidemiologic study subjects. However, the multiple comparisons challenges of relating high-dimensional phenotypes to high-dimensional genotypes and interactions are even more daunting than for regular GWA studies. Alternatively, stand-alone studies or external databases can be used to construct prior covariates to inform G×E analyses of epidemiologic studies. For example, GWA data on immunologic markers for a challenge study of allergen and diesel exhaust particles are being used to define a set of immunologic covariates associated with each SNP as priors in a hierarchical model for a GWA study of asthma. Associations of genome-wide expression with genome-wide SNPs[143] could be used in a similar manner, and could be even more promising for G×E interactions if based on expression studies conducted under a range of environmental conditions.

Next-generation sequencing and rare variants in a G×E context. Increasing attention is being paid to the possibility that rare variants might account for at least some of the missing heritability[144]. Next-generation sequencing methods are making it feasible to sequence portions of the genome identified through a GWA study in a subset of study subjects. Until it becomes possible to obtain and manage genome-wide sequence information on the massive sample sizes that would be required to discover associations with rare variants directly, some form of informative sampling will be required. For example, one might sequence a subsample of cases and controls — stratified by associated SNPs in a given region, family history and environmental factors — to discover novel variants in the region, and a joint analysis could be carried out on the subsample and the main study data[94,145]. The imminent availability of the 1000 Genomes Project[146] data will doubtless have a profound effect on the design of such studies.

Public health and personal medicine implications. Insights from G×E interactions could have important policy implications for environmental health standards[147], the targeting of interventions[148] and treatment selection[149] (BOX 2). For example, the Clean Air Act directs the US Environmental Protection Agency to set standards to protect the most sensitive, including genetically susceptible individuals[150], although it has been argued that public health interventions aimed at the whole population may be more effective[151]. As another example, suppose the joint effect of mutations in *BRCA1* and/or *BRCA2* in combination with RT in an individual was multiplicative; then even if the radiation effect in mutation carriers alone was not statistically significant or the joint effect was not significantly greater than additive, it would be misleading to conclude that RT was no more dangerous for carriers than for non-carriers, as carriers have a much higher baseline risk[152]. Because any statement about interaction is necessarily scale dependent (BOX 1), it is essential that claims about the presence or absence of an interaction make clear whether it is a departure from an additive or multiplicative model on a scale of absolute or attributable risk, odds, underlying liability or some other scale that is being discussed. Unfortunately, the translation of scientific understanding about G×E interactions into risk assessment and prevention policies has so far been limited[153].

Conclusions

The current enthusiasm for studying genetic associations with disease, recently enhanced by the advent of GWA studies, has tended to overshadow the important role of environmental factors and G×E interactions. Although these are much more difficult to study than purely genetic associations due to the need for careful collection of exposure data and rigorous study designs, standard epidemiologic designs can be used, and several recently developed variants of them can enhance power. Nevertheless, large consortia are likely to be needed to fully explore G×E interactions, and such efforts will need to consider these principles and harmonization across studies. The use of powerful pathway-based methods that leverage external biological knowledge can further enhance power and insights.

Latent variable models
A model involving one or more unobservable intermediate variables that represent the pathway connecting a cause (for example, exposures and genotypes) to an effect (for example, disease). Identifying the pathways typically requires the use of surrogates for the latent variables (for example, biomarkers) in addition to the observable cause and effect variables.

1000 Genomes Project
A large-scale effort to obtain and catalogue the full genome-wide DNA sequence of 1,000 individuals selected from a range of races.

© 2010 Macmillan Publishers Limited. All rights reserved

1. Le Marchand, L. The predominance of the environment over genes in cancer causation: implications for genetic epidemiology. *Cancer Epidemiol. Biomarkers Prev.* **14**, 1037–1039 (2005).
2. Le Marchand, L. & Wilkens, L. R. Design considerations for genomic association studies: importance of gene–environment interactions. *Cancer Epidemiol. Biomarkers Prev.* **17**, 263–267 (2008).
3. Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. & Gauderman, W. J. Exploiting gene–environment interaction to detect genetic associations. *Hum. Hered.* **63**, 111–119 (2007).
4. Hunter, D. J. Gene–environment interactions in human diseases. *Nature Rev. Genet.* **6**, 287–298 (2005).
An excellent Review of the basic principles of epidemiological study designs for G×E interactions in the pre-GWA studies era. Among other insights, the author argues that G×E findings can ‘point the finger’ towards the causal constituent of a complex mixture.
5. Greene, C. S., Penrod, N. M., Williams, S. M. & Moore, J. H. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS ONE* **4**, e5639 (2009).
6. Ioannidis, J. P. Non-replication and inconsistency in the genome-wide association setting. *Hum. Hered.* **64**, 203–213 (2007).
7. Thomas, D. Methods for investigating gene–environment interactions in candidate pathway and genome-wide association studies. *Annu. Rev. Public Health* 4 Jan 2010 (doi:10.1146/annurev.publhealth.012809.103619).
8. Cordell, H. J. Detecting gene–gene interactions that underlie human diseases. *Nature Rev. Genet.* **10**, 392–404 (2009).
9. Holmans, P. *et al.* Gene Ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.* **85**, 13–24 (2009).
10. Sebastiani, P., Ramoni, M. F., Nolan, V., Baldwin, C. T. & Steinberg, M. H. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nature Genet.* **37**, 435–440 (2005).
11. Khoury, M. J. & Wacholder, S. Invited commentary: from genome-wide association studies to gene–environment-wide interaction studies — challenges and opportunities. *Am. J. Epidemiol.* **169**, 227–230 (2009).
12. Thomas, D. C. Exposure–time–response relationships with applications to cancer epidemiology. *Ann. Rev. Public Health* **9**, 451–482 (1988).
13. Thomas, D. C., Stram, D. & Dwyer, J. Exposure measurement error: influence on exposure–disease relationships and methods of correction. *Ann. Rev. Public Health* **14**, 69–93 (1993).
14. Lobach, I., Carroll, R. J., Spinka, C., Gail, M. H. & Chatterjee, N. Haplotype-based regression analysis and inference of case–control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics* **64**, 673–684 (2008).
15. Wong, M. Y., Day, N. E., Luan, J. A. & Wareham, N. J. Estimation of magnitude in gene–environment interactions in the presence of measurement error. *Stat. Med.* **23**, 987–998 (2004).
16. Smith, P. G. & Day, N. E. The design of case–control studies: the influence of confounding and interaction effects. *Int. J. Epidemiol.* **13**, 356–365 (1984).
17. Gauderman, W. J. Sample size requirements for matched case–control studies of gene–environment interaction. *Stat. Med.* **21**, 35–50 (2002).
This paper describes a general approach to sample size and power calculations for G×E studies and the capabilities of the freely available Quanto program for this purpose.
18. Garcia-Closas, M. & Lubin, J. H. Power and sample size calculations in case–control studies of gene–environment interactions: comments on different approaches. *Am. J. Epidemiol.* **149**, 689–692 (1999).
19. Burton, P. R. *et al.* Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. *Int. J. Epidemiol.* **38**, 263–273 (2009).
20. Ioannidis, J. P., Trikalinos, T. A. & Khoury, M. J. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am. J. Epidemiol.* **164**, 609–614 (2006).
21. Matullo, G., Berwick, M. & Vineis, P. Gene–environment interactions: how many false positives? *J. Natl Cancer Inst.* **97**, 550–551 (2005).
22. Clayton, D. & McKeigue, P. M. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* **358**, 1356–1360 (2001).
This paper takes a critical look at the current enthusiasm for G×E interactions, particularly in the context of large biobanks. The authors argue for case–control studies over cohort studies and for relying on case-only methods for detecting G×E interactions; however, they question whether genes involved in interactions might not more easily be discovered on the basis of the marginal associations they induce.
23. Moore, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* **56**, 73–82 (2003).
The creator of the MDR algorithm for identifying higher-order interactions gives a spirited argument in support of the notion that many such effects would be overlooked by limiting attention to factors showing significant main effects.
24. Moore, J. H. & Williams, S. M. Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* **85**, 309–320 (2009).
25. Yang, Q. & Khoury, M. J. Evolving methods in genetic epidemiology. III. Gene–environment interaction in epidemiologic research. *Epidemiol. Rev*. **19**, 33–43 (1997).
Another excellent review of study design principles for G×E interactions, covering a broad range of designs.
26. Manolio, T. A., Bailey-Wilson, J. E. & Collins, F. S. Genes, environment and the value of prospective cohort studies. *Nature Rev. Genet.* **7**, 812–820 (2006).
27. Andrieu, N. & Goldstein, A. M. Epidemiologic and genetic approaches in the study of gene–environment interaction: an overview of available methods. *Epidemiol. Rev.* **20**, 137–147 (1998).
28. Piegorsch, W., Weinberg, C. & Taylor, J. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. *Stat. Med*. **13**, 153–162 (1994).
The paper that introduced the case-only design for testing G×E interactions.
29. Caporaso, N. *et al.* Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS ONE* **4**, e4653 (2009).
30. Thorgeirsson, T. E. *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638–642 (2008).
31. Thomas, D. C. Case–parents design for gene–environment interaction by Schaid. *Genet. Epidemiol.* **19**, 461–463 (2000).
32. Broeks, A. *et al.* Identification of women with an increased risk of developing radiation-induced breast cancer: a case only study. *Breast Cancer Res.* **9**, R26 (2007).
33. Albert, P. S., Ratnasinghe, D., Tangrea, J. & Wacholder, S. Limitations of the case-only design for identifying gene–environment interactions. *Am. J. Epidemiol.* **154**, 687–693 (2001).
34. Mukherjee, B. *et al.* Tests for gene–environment interaction from case–control data: a novel study of type I error, power and designs. *Genet. Epidemiol*. **32**, 615–626 (2008).
35. Li, D. & Conti, D. V. Detecting gene–environment interactions using a combined case-only and case–control approach. *Am. J. Epidemiol.* **169**, 497–504 (2009).
36. Schaid, D. Case–parents design for gene–environment interaction. *Genet. Epidemiol.* **16**, 261–273 (1999).
This paper introduced the transmission-disequilibrium test stratified by the case’s exposure as a method of testing for G×E interactions that is robust to population G–E association.
37. Gauderman, W. J., Witte, J. S. & Thomas, D. C. Family-based association studies. *J. Natl Cancer Inst. Monogr.* **26**, 31–37 (1999).
38. Laird, N. M. & Lange, C. Family-based designs in the age of large-scale gene-association studies. *Nature Genet.* **7**, 385–394 (2006).
A review of the various family-based designs for testing genetic main effects in the context of GWA studies.
39. Cui, J. S. *et al.* Regressive logistic and proportional hazards disease models for within-family analyses of measured genotypes, with application to a CYP17 polymorphism and breast cancer. *Genet. Epidemiol.* **24**, 161–172 (2003).
40. Boomsma, D., Busjahn, A. & Peltonen, L. Classical twin studies and beyond. *Nature Rev. Genet.* **3**, 872–882 (2002).
41. Andrieu, N. & Demenais, F. Interactions between genetic and reproductive factors in breast cancer risk in a French family sample. *Am. J. Hum. Genet.* **61**, 678–690 (1997).
42. Gauderman, W. J. & Faucett, C. L. Detection of gene–environment interactions in joint segregation and linkage analysis. *Am. J. Hum. Genet.* **61**, 1189–1199 (1997).
43. Gauderman, W. J. & Siegmund, K. D. Gene–environment interaction and affected sib pair linkage analysis. *Hum. Hered*. **52**, 34–46 (2001).
44. Schaid, D. J., Olson, J. M., Gauderman, W. J. & Elston, R. C. Regression models for linkage: issues of traits, covariates, heterogeneity, and interaction. *Hum. Hered.* **55**, 86–96 (2003).
45. White, J. E. A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am. J. Epidemiol.* **115**, 119–128 (1982).
The paper that first introduced the idea of two-stage sampling in the epidemiologic context.
46. Breslow, N. E. & Chatterjee, N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Appl. Stat.* **48**, 457–468 (1999).
Arguably the most accessible summary of a major series of papers on the design and analysis of two-phase case–control studies.
47. Li, R. *et al.* Glutathione *S*-transferase genotype as a susceptibility factor in smoking-related coronary heart disease. *Atherosclerosis* **149**, 451–462 (2000).
48. Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E. & Kulich, M. Using the whole cohort in the analysis of case–cohort data. *Am. J. Epidemiol.* **169**, 1398–1405 (2009).
An important contribution to the literature on two-phase case–control studies that emphasizes the value added by exploiting the information available on the entire cohort that is not used in standard analysis methods.
49. Bernstein, J. L. *et al.* Study design: evaluating gene–environment interactions in the etiology of breast cancer — the WECARE study. *Breast Cancer Res.* **6**, R199–R214 (2004).
This paper provides an overview of the design of the WECARE study, giving particular attention to the power gained from using the counter-matched design when testing for gene–radiation interactions.
50. Langholz, B. & Goldstein, L. Risk set sampling in epidemiologic cohort studies. *Stat. Sci.* **11**, 35–53 (1996).
This paper provides a non-technical discussion of counter-matching and other cohort sampling designs, with numerous examples of applications for epidemiologic studies.
51. Andrieu, N., Goldstein, A. M., Thomas, D. C. & Langholz, B. Counter-matching in studies of gene–environment interaction: efficiency and feasibility. *Am. J. Epidemiol.* **153**, 265–274 (2001).
52. Gilliland, F. D., McConnell, R., Peters, J. & Gong, H. Jr. A theoretical basis for investigating ambient air pollution and children’s respiratory health. *Environ. Health Perspect.* **107**, 403–407 (1999).
This paper provides a superb overview of the biological rationale for focusing studies of air pollution and respiratory disease on genes and environmental modifiers involved in oxidative stress and inflammatory pathways.
53. Hoh, J., Wille, A. & Ott, J. Trimming, weighting, and grouping SNPs in human case–control association studies. *Genome Res*. **11**, 2115–2119 (2001).
54. McKinney, B. A., Reif, D. M., Ritchie, M. D. & Moore, J. H. Machine learning for detecting gene–gene interactions: a review. *Appl. Bioinformatics* **5**, 77–88 (2006).
55. Moore, J. H. & Williams, S. M. Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* **85**, 309–320 (2009).
56. Ritchie, M. D. & Motsinger, A. A. Multifactor dimensionality reduction for detecting gene–gene and gene–environment interactions in pharmacogenomics studies. *Pharmacogenomics* **6**, 823–834 (2005).
57. Le Marchand, L. *et al.* Combined effects of well-done red meat, smoking, and rapid *N*-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectal cancer risk. *Cancer Epidemiol. Biomarkers Prev.* **10**, 1259–1266 (2001).
A classic example of an interaction involving two genes and two exposures for which none of the constituent lower-order main effects or interactions is significant.

© 2010 Macmillan Publishers Limited. All rights reserved

58. Vineis, P. *et al.* Current smoking, occupation, *N*-acetyltransferase-2 and bladder cancer: a pooled analysis of genotype-based studies. *Cancer Epidemiol. Biomarkers Prev.* **10**, 1249–1252 (2001).
59. Thomas, D. C. *et al.* Approaches to complex pathways in molecular epidemiology: summary of an AACR special conference. *Cancer Res.* **68**, 10028–10030 (2008).
60. Thomas, D. C. The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiol. Biomarkers Prev.* **14**, 557–559 (2005).
61. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
62. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81**, 1278–1283 (2007).
63. Hong, M. G., Pawitan, Y., Magnusson, P. K. & Prince, J. A. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet.* **126**, 289–301 (2009).
64. Chasman, D. I. On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet. Epidemiol.* **32**, 658–668 (2008).
This paper provides a clear discussion of the use of GSEA as a way of prioritizing hits from a GWA study and interpreting the ensemble of SNP associations in relation to pathways.
65. Aragaki, C. C., Greenland, S., Probst-Hensch, N. & Haile, R. W. Hierarchical modeling of gene–environment interactions: estimating *NAT2* genotype-specific dietary effects on adenomatous polyps. *Cancer Epidemiol. Biomarkers Prev.* **6**, 307–314 (1997).
66. Wakefield, J., De Vocht, F. & Hung, R. J. Bayesian mixture modeling of gene–environment and gene–gene interactions. *Genet. Epidemiol.* **34**, 16–25 (2010).
67. Hung, R. J. *et al.* Inherited predisposition of lung cancer: a hierarchical modeling approach to DNA repair and cell cycle control pathways. *Cancer Epidemiol. Biomarkers Prev.* **16**, 2736–2744 (2007).
68. Hung, R. J. *et al.* Using hierarchical modeling in genetic association studies with multiple markers: application to a case–control study of bladder cancer. *Cancer Epidemiol. Biomarkers Prev.* **13**, 1013–1021 (2004).
One of the first examples of the use of hierarchical modelling for the study of G×E interactions. A set of pathway indicator variables are used as prior covariates to classify specific combinations of genes and environmental exposures.
69. Conti, D. V. *et al.* in *Phenotypes and Endophenotypes: Foundations for Genetic Studies of Nicotine Use and Dependence* (ed. Swan, G. E.) 539–584 (NCI Tobacco Control Monographs, Bethesda, Maryland, 2009).
70. Wang, L. & Weinshilboum, R. M. Pharmacogenomics: candidate gene identification, functional validation and mechanisms. *Hum. Mol. Genet.* **17**, R174–R179 (2008).
71. Rebbeck, T. R., Spitz, M. & Wu, X. Assessing the function of genetic variants in candidate gene association studies. *Nature Rev. Genet.* **5**, 589–597 (2004).
An excellent discussion of ways of interpreting candidate-gene associations in relation to biological function. The functions are inferred from various external sources of information or from programs for computing the predicted function of polymorphisms.
72. Ulrich, C. M. *et al.* Mathematical modeling of folate metabolism: predicted effects of genetic polymorphisms on mechanisms and biomarkers relevant to carcinogenesis. *Cancer Epidemiol. Biomarkers Prev.* **17**, 1822–1831 (2008).
One of a long series of papers on mathematical modelling of the folate pathway. This article focuses specifically on the use of the authors' model to predict the effects of variation in metabolic rate parameters for polymorphisms in specific genes on various outcomes, such as homocysteine concentration or DNA methylation reactions.
73. Thomas, D. C. *et al.* Use of pathway information in molecular epidemiology. *Hum. Genomics* **4**, 21–42 (2010).
74. Armitage, P. & Doll, R. The age distribution of cancer and a multistage theory of carcinogenesis. *Br. J. Cancer* **8**, 1–12 (1954).
75. Moolgavkar, S. H. & Knudson, A. G. Jr. Mutation and cancer: a model for human carcinogenesis. *J. Natl Cancer Inst.* **66**, 1037–1052 (1981).
76. Racine-Poon, A. & Wakefield, J. Statistical methods for population pharmacokinetic modelling. *Stat. Methods Med. Res.* **7**, 63–84 (1998).
77. Clewell, H. J., Andersen, M. E. & Barton, H. A. A consistent approach for the application of pharmacokinetic modeling in cancer and noncancer risk assessment. *Environ. Health Persp.* **110**, 85–93 (2002).
78. Bois, F. Y. Applications of population approaches in toxicology. *Toxicol. Lett.* **120**, 385–394 (2001).
79. Nijhout, H. F., Reed, M. C. & Ulrich, C. M. Mathematical models of folate-mediated one-carbon metabolism. *Vitam. Horm.* **79**, 45–82 (2008).
80. Bergman, R. N. *et al.* Minimal model-based insulin sensitivity has greater heritability and a different genetic basis than homeostasis model assessment or fasting insulin. *Diabetes* **52**, 2168–2174 (2003).
81. Cascorbi, I. Genetic basis of toxic reactions to drugs and chemicals. *Toxicol. Lett.* **162**, 16–28 (2006).
82. Cortessis, V. & Thomas, D. C. in *Mechanistic Considerations in the Molecular Epidemiology of Cancer* (eds Bird, P., Boffetta, P., Buffler, P. & Rice, J.) 127–150 (IARC Scientific Publications, Lyon, France, 2003).
83. Thomas, D. C. Multistage sampling for latent variable models. *Lifetime Data Anal.* **13**, 565–581 (2007).
84. Didelez, V. & Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.* **16**, 309–330 (2007).
85. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
86. Greenland, S. An introduction to instrumental variables for epidemiologists. *Int. J. Epidemiol.* **29**, 722–729 (2000).
87. Dai, J. Y., LeBlanc, M. & Kooperberg, C. Semiparametric estimation exploiting covariate independence in two-phase randomized trials. *Biometrics* **65**, 178–187 (2009).
88. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
89. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
90. Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E. & Begg, C. B. Two-stage designs for gene–disease association studies. *Biometrics* **58**, 163–170 (2002).
91. Wang, H., Thomas, D. C., Pe'er, I. & Stram, D. O. Optimal two-stage genotyping designs for genome-wide association scans. *Genet. Epidemiol.* **30**, 356–368 (2006).
92. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Optimal designs for two-stage genome-wide association studies. *Genet. Epidemiol.* **31**, 776–788 (2007).
93. Elston, R. C., Lin, D. & Zheng, G. Multistage sampling for genetic studies. *Annu. Rev. Genomics Hum. Genet.* **8**, 327–342 (2007).
94. Thomas, D. C. *et al.* Methodological issues in multistage genome-wide association studies. *Stat. Sci.* Preprint at http://www.imstat.org/sts/future_papers.html (2009).
95. Kooperberg, C. & Leblanc, M. Increasing the power of identifying gene × gene interactions in genome-wide association studies. *Genet. Epidemiol.* **32**, 255–263 (2008).
96. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.* **37**, 413–417 (2005).
97. Evans, D. M., Marchini, J., Morris, A. P. & Cardon, L. R. Two-stage two-locus models in genome-wide association. *PLoS Genet.* **2**, e157 (2006).
98. Umbach, D. M. & Weinberg, C. R. Designing and analysing case–control studies to exploit independence of genotype and exposure. *Stat. Med.* **16**, 1731–1743 (1997).
99. Murcray, C. E., Lewinger, J. P. & Gauderman, W. J. Gene–environment interaction in genome-wide association studies. *Am. J. Epidemiol.* **169**, 219–226 (2009).
100. Pearson, J. V. *et al.* Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. *Am. J. Hum. Genet.* **80**, 126–139 (2007).
101. Craig, D. W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods* **5**, 887–893 (2008).
102. Sham, P., Bader, J. S., Craig, I., O'Donovan, M. & Owen, M. DNA pooling: a tool for large-scale association studies. *Nature Rev. Genet.* **3**, 862–871 (2002).
103. Cantor, R. M., Lange, K. & Sinsheimer, J. S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* **86**, 6–22 (2010).
104. Roeder, K., Devlin, B. & Wasserman, L. Improving power in genome-wide association studies: weights tip the scale. *Genet. Epidemiol.* **31**, 741–747 (2007).
105. Whittemore, A. S. A Bayesian false discovery rate for multiple testing. *J. Appl. Stat.* **34**, 1–9 (2007).
106. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
107. Wakefield, J. Reporting and interpretation in genome-wide association studies. *Int. J. Epidemiol.* **37**, 641–653 (2008).
108. Datta, S. Empirical Bayes screening of many *p*-values with applications to microarray studies. *Bioinformatics* **21**, 1987–1994 (2005).
109. Chen, G. K. & Witte, J. S. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am. J. Hum. Genet.* **81**, 397–404 (2007).
110. Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J. & Thomas, D. C. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidemiol.* **31**, 871–882 (2007).
111. Binder, H. & Schumacher, M. Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics* **10**, 18 (2009).
112. Holden, M., Deng, S., Wojnowski, L. & Kulle, B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* **24**, 2784–2785 (2008).
113. Elbers, C. C. *et al.* Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet. Epidemiol.* **33**, 419–431 (2009).
114. Baranzini, S. E. *et al.* Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.* **18**, 2078–2090 (2009).
115. Torkamani, A., Topol, E. J. & Schork, N. J. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* **92**, 265–272 (2008).
116. Lesnick, T. G. *et al.* A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet.* **3**, e98 (2007).
117. Thomas, P. D. *et al.* A systems biology network model for genetic association studies of nicotine addiction and treatment. *Pharmacogenet. Genomics* **19**, 538–551 (2009).
118. Gieger, C. *et al.* Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* **4**, e1000282 (2008).
119. Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science* **303**, 799–805 (2004).
An important paper that popularized the use of Bayesian network analysis for the reconstruction of gene networks from gene co-expression data.
120. Ramoni, R. B., Saccone, N. L., Hatsukami, D. K., Bierut, L. J. & Ramoni, M. F. A Testable prognostic model of nicotine dependence. *J. Neurogenet.* **23**, 283–292 (2009).
121. Ferrazzi, F., Sebastiani, P., Ramoni, M. F. & Bellazzi, R. Bayesian approaches to reverse engineer cellular systems: a simulation study on nonlinear Gaussian networks. *BMC Bioinformatics* **8**, S2 (2007).
122. Kohler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).
123. Koch, L. G. & Britton, S. L. Development of animal models to test the fundamental basis of gene–environment interactions. *Obesity (Silver Spring)* **16**, S28–S32 (2008).
124. Gilliland, F. D., Li, Y. F., Saxon, A. & Diaz-Sanchez, D. Effect of glutathione-*S*-transferase M1 and P1 genotypes on xenobiotic enhancement of allergic responses: randomised, placebo-controlled crossover study. *Lancet* **363**, 119–125 (2004).
An excellent example of the use of experimental designs for investigating G×E interactions, in this case a randomized crossover challenge study of immunologic responses to diesel exhaust particles in allergic subjects.

© 2010 Macmillan Publishers Limited. All rights reserved

125. Thomas, D. C. & Conti, D. V. Two stage genetic association studies. in *Encyclopedia of Clinical Trials* (eds D'Agostino, R., Sullivan, L. & Massaro, J.) (Wiley, New York, 2007).

126. Israel, E. *et al.* Use of regularly scheduled albuterol treatment in asthma: genotype-stratified, randomised, placebo-controlled cross-over trial. *Lancet* **364**, 1505–1512 (2004).

127. Davis, B. R. *et al.* Imputing gene–treatment interactions when the genotype distribution is unknown using case-only and putative placebo analyses — a new method for the Genetics of Hypertension Associated Treatment (GenHAT) study. *Stat. Med.* **23**, 2413–2427 (2004).

128. Vittinghoff, E. & Bauer, D. C. Case-only analysis of treatment–covariate interactions in clinical trials. *Biometrics* **62**, 769–776 (2006).

129. Lin, B. K. *et al.* Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database. *Am. J. Epidemiol.* **164**, 1–4 (2006).

130. Khoury, M. J. & Little, J. Human genome epidemiologic reviews: the beginning of something HuGE. *Am. J. Epidemiol.* **151**, 2–3 (2000).

131. Yesupriya, A. *et al.* Reporting of human genome epidemiology (HuGE) association studies: an empirical assessment. *BMC Med. Res. Methodol.* **8**, 31 (2008).

132. Jensen, L. J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Rev. Genet.* **7**, 119–129 (2006).

133. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).

134. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* **34**, D322–D326 (2006).

135. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2008).

136. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).

137. Miller, R. L. & Ho, S. M. Environmental epigenetics and asthma: current concepts and call for studies. *Am. J. Respir. Crit. Care Med.* **177**, 567–573 (2008).

138. Salk, J. J., Fox, E. J. & Loeb, L. A. Mutational heterogeneity in human cancers: origin and consequences. *Annu. Rev. Pathol.* **5**, 51–75 (2010).

139. Zeisel, S. H. Epigenetic mechanisms for nutrition determinants of later health outcomes. *Am. J. Clin. Nutr.* **89**, 1488S–1493S (2009).

140. Perera, F. *et al.* Relation of DNA methylation of 5′-CpG island of *ACSL3* to transplacental exposure to airborne polycyclic aromatic hydrocarbons and childhood asthma. *PLoS ONE* **4**, e4488 (2009).

141. Baccarelli, A. *et al.* Rapid DNA methylation changes after exposure to traffic particles. *Am. J. Respir. Crit. Care Med.* **179**, 572–578 (2009).

142. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl Acad. Sci. USA* **102**, 10604–10609 (2005).

143. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).

144. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

145. Zhu, X., Feng, T., Li, Y., Lu, Q. & Elston, R. C. Detecting rare variants for complex traits using family and unrelated data. *Genet. Epidemiol.* **34**, 171–187 (2010).

146. Siva, N. 1000 Genomes project. *Nature Biotech.* **26**, 256 (2008).

147. Cullen, A. C., Corrales, M. A., Kramer, C. B. & Faustman, E. M. The application of genetic information for regulatory standard setting under the clean air act: a decision-analytic approach. *Risk Anal.* **28**, 877–890 (2008).

148. Shostak, S. Locating gene–environment interaction: at the intersections of genetics and public health. *Soc. Sci. Med.* **56**, 2327–2342 (2003).

149. Need, A. C., Motulsky, A. G. & Goldstein, D. B. Priorities and standards in pharmacogenetic research. *Nature Genet.* **37**, 671–681 (2005).

150. Lave, L. B. & Omenn, G. S. *Clearing The Air: Reforming The Clean Air Act* (Brookings Institution, Washington, DC, 1981).

151. Rose, G. *The Strategy Of Preventive Medicine* (Oxford Univ. Press, 1992).

152. Bernstein, J. L. *et al.* Radiation-induced second primary breast cancer and *BRCA1* and *BRCA2* mutation carrier status: a report from the WECARE Study. *J. Natl Cancer Inst.* (in the press).

153. Perera, F. P. Molecular epidemiology: on the path to prevention? *J. Natl Cancer Inst.* **92**, 602–612 (2000).

154. Feng, D. *et al.* Platelet glycoprotein IIIa *Pl*^{A} polymorphism, fibrinogen, and platelet aggregability: The Framingham Heart Study. *Circulation* **104**, 140–144 (2001).

155. He, C., Tamimi, R. M., Hankinson, S. E., Hunter, D. J. & Han, J. A prospective study of genetic polymorphism in MPO, antioxidant status, and breast cancer risk. *Breast Cancer Res. Treat.* **113**, 585–594 (2009).

156. Bureau, A., Diallo, M. S., Ordovas, J. M. & Cupples, L. A. Estimating interaction between genetic and environmental risk factors: efficiency of sampling designs within a cohort. *Epidemiology* **19**, 83–93 (2008).

157. Jugessur, A. *et al.* Cleft palate, transforming growth factor alpha gene variants, and maternal exposures: assessing gene–environment interactions in case–parent triads. *Genet. Epidemiol.* **25**, 367–374 (2003).

158. Mayer, E. J. *et al.* Genetic and environmental influences on insulin levels and the insulin resistance syndrome: an analysis of women twins. *Am. J. Epidemiol.* **143**, 323–332 (1996).

159. Bernstein, J. L. *et al.* Radiation exposure, the *ATM* gene, and risk of bilateral breast cancer in the WECARE study. *J. Natl Cancer Inst.* (in the press).

160. Gilliland, F. D. *et al.* Effects of glutathione *S*-transferase M1, maternal smoking during pregnancy, and environmental tobacco smoke on asthma and wheezing in children. *Am. J. Respir. Crit. Care Med.* **166**, 457–463 (2002).

161. Martinez, F. D. Gene–environment interactions in asthma: with apologies to William of Ockham. *Proc. Am. Thorac. Soc.* **4**, 26–31 (2007).

162. Gianfagna, F., De Feo, E., van Duijn, C. M., Ricciardi, G. & Boccia, S. A systematic review of meta-analyses on gene polymorphisms and gastric cancer risk. *Curr. Genomics* **9**, 361–374 (2008).

163. Siemiatycki, J. & Thomas, D. C. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int. J. Epidemiol.* **10**, 383–387 (1981).

164. Greenland, S. Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology* **20**, 14–17 (2009).

165. Haldane, J. B. S. *Heredity and Politics* (W. W. Norton, New York, 1938).

166. Ottman, R. An epidemiologic approach to gene–environment interaction. *Genet. Epidemiol.* **7**, 177–185 (1990).
This widely quoted paper was one of the first to offer a classification of different types of G×E interactions, and gives classic examples of each type.

167. Lewontin, R. C. Annotation: the analysis of variance and the analysis of causes. *Am. J. Hum. Genet.* **26**, 400–411 (1974).

168. Garcia-Closas, M. *et al.* *NAT2* slow acetylation, *GSTM1* null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet* **366**, 649–659 (2005).

169. Dearfield, K. L., Benson, W. H., Gallagher, K. & Johnson, J. D. in *Genomics and Environmental Regulation: Science, Ethics, and Law* (eds Sharp, R. R., Marchant, G. E. & Grodsky, J. A.) 25–34 (Johns Hopkins Univ. Press, Baltimore, 2009).

170. Lympany, P. A. *et al.* HLA-DPB polymorphisms: Glu 69 association with sarcoidosis. *Eur. J. Immunogenet.* **23**, 353–359 (1996).

171. Jacobi, C. E., Nagelkerke, N. J., van Houwelingen, J. H. & de Bock, G. H. Breast cancer screening, outside the population-screening program, of women from breast cancer families without proven *BRCA1/BRCA2* mutations: a simulation study. *Cancer Epidemiol. Biomarkers Prev.* **15**, 429–436 (2006).

172. Ulrich, C. M. & Potter, J. D. Folate supplementation: too much of a good thing? *Cancer Epidemiol. Biomarkers Prev.* **15**, 189–193 (2006).

Acknowledgements

Supported in part by grants 5P30 ES007043, 1U01 ES15090 and 1R01 ES016813 from the US National Institute of Environmental Health Sciences. The author is grateful to D. Conti, W. J. Gauderman, F. Gilliland and R. Haile for many helpful suggestions.

Competing interests statement

The author declares no competing financial interests.

DATABASES

Entrez Gene: http://www.ncbi.nlm.nih.gov/gene
ADRB2 | *ATM* | *BRCA1* | *BRCA2* | *CHEK2* | *CYP1A2* | *GSTM1* | *GSTT1* | *NAT2*

FURTHER INFORMATION

Duncan Thomas's homepage: http://hydra.usc.edu/Thomas
Human Genome Epidemiology Network (HuGENet): http://www.hugenet.org.uk
***Nature Reviews Genetics* series on Genome-wide association studies:** http://www.nature.com/nrg/series/gwas/index.html
POWER: http://dceg.cancer.gov/tools/design/power
Quanto: http://hydra.usc.edu/gxe

SUPPLEMENTARY INFORMATION

See online article: S1 (figure)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2010 Macmillan Publishers Limited. All rights reserved