Hum Genet (2012) 131:1591–1613
DOI 10.1007/s00439-012-1192-0

REVIEW PAPER

Challenges and opportunities in genome-wide environmental interaction (GWEI) studies

Hugues Aschard · Sharon Lutz · Bärbel Maus · Eric J. Duell · Tasha E. Fingerlin · Nilanjan Chatterjee · Peter Kraft · Kristel Van Steen

Received: 1 March 2012 / Accepted: 11 June 2012 / Published online: 4 July 2012

Abstract The interest in performing gene–environment interaction studies has seen a significant increase with the increase of advanced molecular genetics techniques. Practically, it became possible to investigate the role of environmental factors in disease risk and hence to investigate their role as genetic effect modifiers. The understanding that genetics is important in the uptake and metabolism of toxic substances is an example of how genetic profiles can modify important environmental risk factors to disease. Several rationales exist to set up gene–environment interaction studies and the technical challenges related to these studies—when the number of environmental or genetic risk factors is relatively small—has been described before. In the post-genomic era, it is now possible to study thousands of genes and their interaction with the environment. This brings along a whole range of new challenges and opportunities. Despite a continuing effort in developing efficient methods and optimal bioinformatics infrastructures to deal with the available wealth of data, the challenge remains how to best present and analyze genome-wide environmental interaction (GWEI) studies involving multiple genetic and environmental factors. Since GWEIs are performed at the intersection of statistical genetics, bioinformatics and epidemiology, usually similar problems need to be dealt with as for genome-wide association gene–gene interaction studies. However, additional complexities need to be considered which are typical for large-scale epidemiological studies, but are also related to "joining" two heterogeneous types of data in explaining complex disease trait variation or for prediction purposes.

Introduction

Experimental studies in model organisms have provided several evidences of interactions between genes and exposures. For a review about the utility of mouse models in the detection of gene–environment interaction effects and the limitations on their application, we refer to Willis-Owen and Valdar (2009). These animal models may be helpful in suggesting candidate gene–environment interactions, but

S. Lutz and B. Maus contributed equally to this work.

H. Aschard (✉) · P. Kraft
Department of Epidemiology, Harvard School of Public Health,
Boston, MA, USA
e-mail: haschard@hsph.harvard.edu

S. Lutz · T. E. Fingerlin
Departments of Epidemiology & Biostatistics and Informatics,
Colorado School of Public Health, University of Colorado
Anschutz Medical Campus, Aurora, CO, USA

B. Maus · K. Van Steen
Systems and Modeling Unit, Montefiore Institute,
University of Liège, Liège, Belgium

B. Maus · K. Van Steen
Bioinformatics and Modeling, GIGA-R, University of Liège,
Liège, Belgium

E. J. Duell
Unit of Nutrition, Environment and Cancer, Epidemiology
Research Program, Catalan Institute of Oncology (ICO),
Bellvitge Biomedical Research Institute (IDIBELL),
Barcelona, Spain

N. Chatterjee
Division of Cancer Epidemiology and Genetics, National Cancer
Institute, Bethesda, MD, USA

P. Kraft
Department of Biostatistics, Harvard School of Public Health,
Boston, MA, USA

epidemiological studies—although more complicated—are needed if we ever want to have a complete understanding of the genetic architecture of complex human diseases. Most common complex diseases are believed to be the result of the combined effect of genes, environmental factors, and their interactions. Throughout this document, we will use the terms exposure and environment interchangeably.

The term "gene–environment interaction" is often loosely used as referring to the interplay of gene and environment in some way. A first clear reporting of different categories of gene–environment interactions dates back from 1938 as referred to in Smith et al. (2008). Here, we define it via "biological" or "statistical" interaction. A biological gene–environment interaction occurs when one or more genetic and one or more environmental factors participate in the same causal mechanism in the same individual (Rothman et al. 2008; Yang and Khoury 1997). One popular and appealing formal definition of "biological interaction" invokes the sufficient component cause model of causation. In this setting, there is one sufficient component cause that involves both the genetic and environmental exposure (Rothman and Greenland 1998; Tchetgen Tchetgen and VanderWeele 2012) (we note that this definition of "biological interaction" does not imply anything about the biochemical mechanism of how genes and environment combine to cause disease).

In contrast, the statistical interactions, which are typically defined as modifications of the effect on one factor by the levels of the other factor in some underlying scale (Bhattacharjee et al. 2010; Greenland 2009; Siemiatycki and Thomas 1981; Thompson 1991), do not imply any inference about a particular biological mode of action. Statistical interactions can be clustered variously based on the specificity of the underlying statistical models. The common classification distinguishes between "quantitative interaction" and "qualitative interaction". Quantitative interaction refers to the presence of a factor (e.g., an exposure) that modified the magnitude of the effect of a second factor (e.g., a mutation) without changing the direction of the effect. On the other hand qualitative interaction refers to situation where a factor will either cancel or reverse the effect of another factor. For additional details on these definitions, see Clayton (2009) or Thomas (2010a). For example of statistical models of interactions see, for example, Wright et al. (2002) or Dempfle et al. (2008).

Gene–environment interaction effects have been investigated for a wide range of candidate genes and exposures for many complex traits, such as cancer, depression, type 2 diabetes, and asthma (Franks 2011; Hunter 2005; Lesch 2004; Stern et al. 2002; Vercelli 2010; Wu et al. 2011). However, only a handful of the large number of reported statistically significant interactions has been replicated, despite well-powered replication efforts for some influential preliminary reports (Cornelis et al. 2011; Dunn et al. 2011; Risch et al. 2009). The candidate gene interaction literature suffers from many of the same problems that plagued the literature on marginal effects of candidate genes, including small sample sizes and inappropriate (or lack of) adjustment for multiple testing. Moreover, replication in the context of gene–environment interaction effects faces additional challenges, including differences in exposure measurement protocols across studies, differences in the scale of reported gene–environment interaction effects, and differences in the distribution of exposures across studies. The candidate gene interaction literature can therefore only provide limited guidance on the number and size of gene–environment interaction effects expected to truly exist in human populations, although it does suggest that large and pervasive interaction effects are unlikely.

Genome-wide approaches to identify loci involved in gene–environment interactions have just begun to appear in the peer-reviewed literature (Ege et al. 2011; Hamza et al. 2011; Paré et al. 2010). For example, Ege et al. (2011) recently completed a genome-wide environment interaction (GWEI) study for childhood asthma and farming exposures in the context of GABRIEL (A Multidisciplinary Study to Identify the Genetic and Environmental Causes of Asthma in the European Community). Although this study was well powered to detect gene–environment interactions for common alleles, no interactions were statistically significant, not even those interactions involving genetic markers in genes previously reported to show interactions (Ober and Vercelli 2011). Developing methods to overcome the conceptual, technical, and methodological hurdles GWEI studies involve is the focus of much ongoing methodological work.

Gene–environment interaction at the age of genome-wide data has been recently discussed in several reviews (Dempfle et al. 2008; Hunter 2005; Khoury and Wacholder 2009; Thomas 2010a, b). In this review, we focus on strategies and methodological aspects of genome-wide association study of gene–environment interactions. In particular, we provide an overview of possible analytical choices in relation to researchers' aims and beliefs. Simply stated, what are the main advantages and disadvantages of the existing approaches based on the goal: identifying new genetic variants involved in gene–environment interactions, identifying gene–environment interaction per se or screening for potential interactions without testing?

The quest for gene–environment interactions

The interest in studying the combine effect of genes and environmental factors in the etiology of common multifactorial disease has grown up in parallel with the study of their genetic component only. In the past 10 years large investments have been made trying to elucidate some of these mechanisms. The UK Medical Research Council, the Wellcome Trust, and the Department of Health, for example, have launched in 2002 the BioBank UK study, a prospective cohort study of 500,000 individuals, which attempts to integrate the genetic and environmental components of disease risk (Wright et al. 2002). The National Institutes of Health (NIH) has initiated the Genes, Environment, and Health Initiative (GEI). It includes the Gene Environment Association Studies (GENEVA) consortium which was established to facilitate the identification of variations in gene-trait associations related to environmental exposures (Cornelis et al. 2010). More recently, the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the University of California San Francisco have launched a new resource for studying disease, health, and aging. In this project, DNA and exposure to environmental factors are collected for more than 100,000 samples.

Besides pharmacogenomics, which represent a particular (and promising) field of study for gene–environment interaction (Meyer 2000; Wright et al. 2002), there are three common arguments that have been emphasized for searching for the presence of gene–environment interactions in common multifactorial diseases. First, for most of the identified genetic variants in genome-wide association studies (GWAS), the mechanisms through which genetic variants contribute to the associated complex phenotypes remains largely unknown. Second, the predictive potential of common genetic variants that have been extensively studied in genome-wide scan appears to be limited (Gibson 2010; Visscher et al. 2010; Yang et al. 2010). Third, the common SNPs that have been identified so far only explain a small proportion of the variance of complex traits. Overall, interaction effects with environmental factors are considered one possible key to a better understanding of the genetic architecture of complex traits (Manolio and Collins 2007; Zuk et al. 2012). Gene–environment interactions might also be further translated into improvement in our ability to predict disease risk and be of utility for various personalized medicine applications, such as targeting individuals that may need costly intervention (Rothman et al. 1980).

However, this ideal picture needs to be balanced by our current knowledge of statistical interaction effect in epidemiology. First, it is notoriously difficult to make inference regarding biological mechanisms from epidemiologic data, and interaction reflects a level of complexity that makes such inference even harder (Clayton 2009; Greenland 2009; Siemiatycki and Thomas 1981; Thompson 1991). Second, interactions are unlikely to dramatically improve risk prediction if they have only moderate effects or if the number of interactions is low (Aschard et al. 2012). Third, the identification of any interaction effect is recognized as an extremely challenging task and the lack of discoveries clearly confirms this issue. Hence a reasonable consensus is that gene–environment interaction studies may at least help in the discovery of new genetic variants and new environmental risk factors, (Gauderman and Thomas 2001; Kraft et al. 2007; Manolio and Collins 2007), which remains an important step toward our understanding of complex diseases.

Our ability to attain some of these goal increases with the growing number of rich heterogeneous data resources, with data available on genetics, family history, physical and behavioral characteristics, life-style, intra-individual changes over time, etc. However, it also comes with some caveats. Despite the fact that these data allow the investigation of more complex, possible non-linear relationships between genetic and non-genetic factors, it remains the question whether the toolbox that is available to date contains sufficiently refined tools and methodologies to be applied in a genome-wide context. Compared with the total number of papers published on gene–environment interactions, GWEI studies only represent a handful of studies (Fig. 1). While we believe gene–environment interaction are more and more studied at the genome-wide scale, the low number of publications may be partially explained by the non-publication of negative results. It may indicate that there is still room for novel approaches and rigorous strategies that can overcome some of the hurdles scientists are facing when performing a GWEI study.

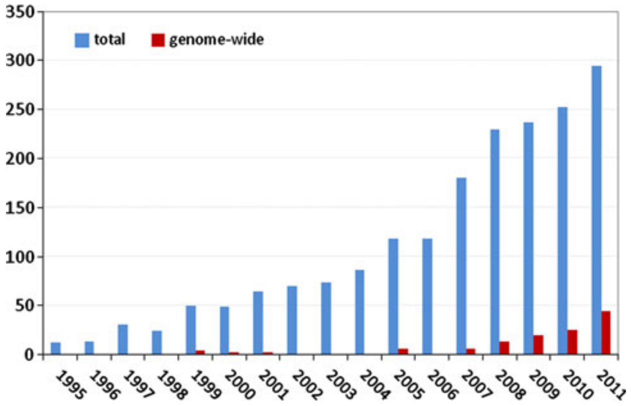

Fig. 1 Number of papers in PubMed with (“gene–environment” or “gene-by-environment” or “gene × environment”) and “interaction” in the title or abstract (in *blue*). Furthermore, the number of papers is shown which additionally to the previous search term also contain (“genome-wide” or genomewide) in the title or abstract (in *red*). It should be noted that this search only retrieves “potential” GWEI studies and that the real numbers of GWEI studies are probably even lower than the reported counts

What are possible complicating factors in GWEI studies?

Confounding

Confounding may occur when independent variables are associated with one another and with the outcome of interest. In epidemiology it refers to a situation when an extraneous variable that cause the phenotype under study is also associated with a predictor of interest that is not causal (i.e., that is not on the "causal pathway" of the phenotype). The existence of confounding variables can make it difficult to establish a clear causal link between the studied predictor and the outcome unless appropriate methods are used to adjust for the effect of the confounders. However, dealing with known confounders is relatively easy. It can be minimized or controlled by a study design or by employing appropriate data analysis methods such as multiple regression or stratification analyses (Demissie and Cupples 2011; Rothman et al. 2008). Dealing with unknown confounders is obviously much trickier, although recent work has shown that unknown confounders of the interacting factors may not necessarily bias the estimation of interaction effect per se (Tchetgen Tchetgen and VanderWeele 2012). It should also be noted that the case-only technique is more likely to be subject to confounding. For example, when analyzing related individuals, family-history, which is related to genetic susceptibility as well as life-style exposures, may create artificial dependencies between a mutation and an exposure. Such confounding effects may invalidate the case-only test while it may be easily handled by using family-data methods (Thomas 2000). Confounding due to latent population substructures, when unintentionally including groups of different ethnicity, is also known to have a larger impact on the validity of the case-only test of interaction than on the case–control interaction test (Wang and Lee 2008).

Exposure measurement error and misclassification

The detection of *G–E* interactions can be severely hampered by unreliability in the assessments of exposures. Measurement challenges for underlying key exposures (e.g., diet, physical activity, air pollution parameters) present important barriers to interaction identification, but equally the assessment of their marginal impact on disease trait (Prentice 2011). Measurement error (or misclassification when explanatory variables in regression models are categorical) is a well-known issue in association studies that can both bias point estimates and generate invalid association. In general, conventional parametric and non-parametric regression techniques are no longer valid when errors in the predictors are expected. Improved study design and methods for corrections have been widely discussed in studies of a single factor (Bashir and Duffy 1995). More recently, attention has been given to the impact of exposure measurement error in *G–E* interaction studies (Carroll et al. 2006; Wong et al. 2004). Despite the fact that various solutions are around to handle measurement error during the statistical analysis (Garcia-Closas et al. 1998, 1999; Lindstrom et al. 2009; Lobach et al. 2011; Thomas 2010b), these methods are not widely used in practice, even for smaller-scaled *G–E* interaction studies. Another consideration about exposure measurement error is that the error structures of environmental exposures may differ across populations and this could have implications for how interactions are detected and interpreted.

In practice, misclassification is usually addressed from two perspectives: (a) how to correct for misclassification in statistical test and (b) how to define the trade-off between sample size and measurement precision to maximize statistical power. The common approach to account for misclassification in statistical test is to use validation studies. It consists in measuring repeatedly a fraction of the sampled subjects with the same error-prone instrument to obtain estimates of misclassification probabilities. Various statistical techniques can be built on this framework. Some of them have been recently described by Zhang et al. (2008) who also introduce simple and practically useful concepts to minimize the biases of all parameters of interest in the presence of both genotyping and exposure misclassification errors. Unfortunately, validation or repeated measurement data that is required to apply such methods in practice are not available in typical studies. When the misclassification issue is considered at the design stage, the perspective is slightly different. Since improving the measurement can be achieved by taking repeated measurements for all individuals (provided the error in repeated measures is uncorrelated), the question is how to balance quantity and quality. Obviously, for a fixed total number of subject evaluations, the use of multiple measurements per subject would result in a halving of sample size. Wong et al. (2003) provide arguments for this strategy by showing that smaller studies with reasonably accurate measurement might be more efficient than larger studies with poor assessment of exposure and outcome when the goal is testing for interaction per se. However, this result does not necessarily hold when the goal is rather to identify genetic variants while allowing for potential interaction effect. In this case, testing for global genetic effect over multiple exposure strata may conserve reasonable power when misclassifications remain low, while the standard test of interaction can suffer a dramatic loss of power (Lindstrom et al. 2009).

Population stratification and population dependencies

Concerns about the widespread of population stratification or the bias it may induce have been raised before. Several approaches to population stratification in main effects GWAS studies are available and commonly in use (Price et al. 2010). Population stratification also becomes an issue in *G–E* interaction studies if subpopulation membership based on genetics is associated with the outcome, the genetic effect, and the environmental exposure. In contrast to GWAS studies, it is less clear how to correct for population stratification and cryptic relatedness in GWEI studies, since strata or degrees of relatedness may be related to the environmental exposure under investigation. It was recently shown that principal component methods, that have been popular for correction of population stratification in GWAS studies, can be used for adjustment of gene–gene or gene–environment dependence due to population stratification in interaction studies (Bhattacharjee et al. 2010).

Alternatively, one can use family-based methods that condition on parental genotypes, which are thought to be robust against population stratification (Laird and Lange 2006). However, recent work by Shi et al. (2011) showed that the standard family-based tests of gene–environment interaction can be biased when the tested genetic variant is not itself the causal variant but a proxy for it (i.e., in linkage disequilibrium with the causal) and the studied exposure does participate in population structure (i.e., when the exposure is correlated with the genotypic strata). They present a solution to correct for such bias when exposure is binary which consists in adjusting for a family-based measure of the exposure distribution. Explicitly they fit a saturated model for the genetic main effect within strata defined by the siblings' exposure profile (exposure need to be collected for an unaffected sibling). Although the empirical extent of the example presented by Shi and colleagues is unknown, there are realistic scenarios where such bias may occur; especially when analyzing recently admixed population such as African–American or Latino (Kraft 2011).

Dynamics of gene–environment interactions

Many exposures change over time and may be prevalent in one population and rare or absent in another. Thus, the amount of population variation in a disease that can be explained by one or more exposures may not be generalized from one population to another or from one time period to another (Pearce 2011). The dynamic "behavior" of an exposure is a function of its prevalence over time in an individual and in a population of interest. The nature of the exposure may also be relevant in terms of *G–E* interaction effects (e.g., the dose and route of exposure, when exposure first or last occurred, or whether exposures were periodic, continuous, intermittent, or single events). Furthermore, there could be critical windows of exposure (etiologically relevant exposure periods), when the exposure is more or less likely to contribute to, or may even have opposing effects on, a disease process. This includes, for example, conception, fetal development, early childhood, and adulthood, before or after the menopause. Several studies have already been successful in identifying such effects (Balansky et al. 2012; Bouzigon et al. 2008; Doherty et al. 2009; Lo et al. 2009). As mentioned earlier, the calendar time period may also be important since many exposures and exposure opportunities change over time (e.g., environmental tobacco smoke, environmental pollution, processed foods, and pharmaceutical drugs).

To the extent that this is possible in ongoing and future prospective cohort studies, exposure should be periodically re-assessed over the course of a study. The ideal design would be a life course approach in which exposure information is collected at different time points throughout an individual's life. Such a study would be cost-prohibitive for most investigators, but very large cohorts of individuals that include extended measurement to a range of exposures and genetic data are now in progress. The aforementioned RPGEH project, for example, includes comprehensive longitudinal health information over long period and will offer the opportunity to explore some of these aspects. Finally, gene-by-'timing of exposure' effects might also be amenable to study in animal model systems (models from conception to death). Such model systems may help to inform the potential critical windows of exposure and relevant mechanisms in humans.

Power and sample size

Perhaps one of the greatest challenges in GWEI studies is that of power (Bookman et al. 2011; Murcray et al. 2011; Thomas 2010a). Inadequate sample sizes give rise to underpowered studies and increase the occurrence of false-positive and false-negative findings. Only a handful of software packages or programs are available to compute sample size and power for *G–E* interaction studies (Dempfle et al. 2008). For a simple interaction model between a single genetic variant and binary or continuous exposure, Murcray et al. (2011) derived the sample size required to achieve 80 % power, for a variety of *G–E* interaction tests, while correcting for multiple testing at the genome-wide level. Their study clearly shows that for moderate to low effects, the required sample size for classical tests is likely to be extremely large, larger than for similar tests of marginal effects with the same amplitude. Obviously, the improved efficiency (increasing power while keeping the same sample size) by using one methodology over another, will highly depend on the mode of interaction. Simulation strategies such as the one developed by Amato et al. (2010), accommodating non-linear interactions, may further help in elucidating the scenario's in which a particular method performs best. Unfortunately, most studies deriving sample size and power calculations in simulated data assume no error in the assessment of genetic factors nor environmental factors, whereas these are known to induce power loss (Garcia-Closas et al. 1999; Tung et al. 2007). It leaves no doubt that there is still room for additional simulation strategies of *G*–*E* interactions, allowing for differential modes of interaction, that are flexible to incorporate some of the aforementioned complicating factors.

Methods

Defining aims and fitting the context

We have compiled a list of papers which define or explore (via simulation or theoretic development) the properties of methods for investigation of gene environment interactions (Table 1). The methods papers listed cover a range of study designs from family-based to case–control to case-only methods. While not exhaustive, the list covers the majority of such research papers published prior to development of this review. In particular, the entries of Table 1 address whether the method is applicable to gene–gene interactions, whether the method is tailored to genome-wide studies or candidate gene studies, and for which type of outcome the method is tailored (i.e., binary, continuous, etc.). While many of the methods can be extended beyond what has currently been described, we limited our categorization to those situations explicitly discussed in the research paper. The table demonstrates the sheer number of methods that are available and illustrates the difficulty in determining which method is appropriate for a given study/situation. For many methods, there is no clear point of comparison or clear choice as to which method is superior.

Naively, any data analysis can be decomposed in three tightly linked cornerstones: (1) the analysis type which is in a one-to-one correspondence with the problem type or research question, (2) the sampling design which aims to maximize the efficiency for a fixed number of individual, and (3) the (statistical) model or methodology which summarizes the (statistical) answer to the research question.

We do not address specifically the measurement type of the variables included, which is related in GWEI studies to traits, genetic markers and exposures. A discussion of the types of genetic markers (e.g., SNPs or CNVs) or measurement scales of exposure variables falls outside the scope of this work. We merely want to highlight that the most commonly used genetic markers used in GWEI studies are SNPs and that the most popular coding is additive, while other type of genetic variations such as CNV (e.g., Karageorgi et al. 2011) or epigenetic markers are barely used. Related to the popularity of the case–control design, traits are often quantified via a binary variable (see also Table 1), although many quantitative traits have also been studied at the genome-wide scale. We discuss below study designs and statistical models that allow handling either binary or quantitative outcome or both.

Cornerstone 1: research problem

Methodological requirements for identifying *G*–*E* interactions are largely driven by the research question and the viewpoint. From a public health perspective, the objective will usually be testing for genetic variant while allowing for interaction or testing for public health interactions (Siemiatycki and Thomas 1981). In such a situation one may use analytic methods making assumptions about the functional form of models and/or effects being modeled and derive an appropriate test to derive effect size estimates and test the hypothesis of interest. In human genetics, two popular analysis types are linkage and association studies. *G*–*E* interaction studies in linkage studies may involve performing exposure stratified analyses (e.g., Colilla et al. 2003) or *G*–*E* interaction testing strategies using sib-pairs (e.g., Dizier et al. (2003) for a review). Here, we will restrict attention to genetic association problems.

It is less clear what test of interaction is most appropriate when the goal of the study is to draw inference about biological mechanism. A significant test for interaction—whether from a multiplicative odds ratio model or additive absolute risk model for disease traits, or from additivity for log-transformed or untransformed continuous traits—need not imply biological interaction, just as biological interaction need not imply statistical interaction (Greenland 2009; Siemiatycki and Thomas 1981; Thompson 1991). The observed distribution of traits across the strata defined by genotype and exposure may be suggestive of underlying biological mechanism, but it is suggestive at most. Formally testing whether a hypothesized null interaction model is contradicted by observed epidemiologic data requires careful mathematical modeling of how the proposed biological mechanism would affect the observed trait distributions—and such modeling will always require untestable assumptions (Thompson 1991).

Table 1 Overview of methods papers for studying of gene–environment interactions

Manuscript	Manuscript type[a]	Marker design[b]	Subject design[c]	Outcome type[d]	Aim[e]	Analysis approach[f]	Population stratification addressed	Applicable to gene–gene interaction
Albrechtsen et al. (2007)	2, 3, 4	1, 2	3	1, 2	1	2	No	Yes
Andrieu and Goldstein (2004)	2, 3	2	2, 5	1	1	1	No	No
Aschard et al. (2011)	2, 3	1, 2	2, 3, 4	1, 2	1	1	No	No
Bureau et al. (2005)	2, 3, 4	2	2	1	1, 5	2	No	Yes
Bůžková et al. (2011)	2, 3	2	2	1, 2	1	1, 2	No	Yes
Cattaert et al. (2010)	2, 3, 4	2	4, 5	2	1, 3, 5	2	No	Yes
Chanda et al. (2009a)	2, 3, 4	2	1, 2	2	1, 3, 5	2	No	Yes
Chanda et al. (2009b)	2, 3, 4	2	2	1, 5	1, 3	2	No	Yes
Chanda et al. (2008)	2, 3, 4	2	2	1, 5	1, 5	2	No	Yes
Chanda et al. (2007)	2, 3, 4	2	2	1, 5	1, 3	2	No	Yes
Chatterjee et al. (2006)	2, 3, 4	2	1, 2	1, 2	1, 2	2	No	Yes
Chatterjee et al. (2005)	2, 3	1, 2	5	1	1, 2	1	Yes	No
Chen et al. (2009b)	2, 3	2	1, 5	1	1, 2	1	Yes	No
Chen et al. (2008)	2, 3	2	2	1	1	1	Yes	No
Chen et al. (2007)	2, 3, 4	2	2	1	1, 2, 5	2	No	Yes
Cheng (2006)	2, 3, 4	1, 2	4	1	1, 2	1	No	Yes
Cordell et al. (2004)	2, 3	1, 2	5	1	1	1, 2	No	Yes
Culverhouse et al. (2004)	2, 3, 4	2	4	2	1, 3, 5	2	No	Yes
Dai et al. (2010)	2,3	1,2	2,3	1,2	1	1	No	No
Efird (2005)	2, 3	2	2	1	1, 2	1	No	No
Fan et al. (2011)	2, 3, 4	2	2	1	1, 3, 5	2	No	Yes
Fardo et al. (2012)	2, 3	2	5	1, 2	1	1	Yes	No
Gauderman et al. (2010)	2, 3	1, 2	5	1	1	2	No	No
Gauderman and Faucett (1997)	2, 3	3	5	2	1	2	No	No
Geneletti et al. (2011)	5	2	4	1	1	1	No	No
Gu et al. (2009)	2, 3, 4	1, 2	4	1	2	1	Yes	No
Hoffmann et al. (2009)	2, 3	1, 2	5	1	1	1	Yes	No
Kazma et al. (2011)	2, 3	1, 2	4	1	1, 2	1	No	No
Kraft et al. (2007)	2, 3	1, 2	2, 4	1, 2	1, 2	1	No	No
Hothorn et al. (2006)	2, 4	3	1, 2	1, 2, 3, 4	1, 3	2	No	Yes
Lake and Laird (2004)	2, 3	1, 2	5	1	1	1	Yes	No
Lee and Chang (2006)	2	2	1, 4	1	1	1	No	No
Li and Conti (2009)	2, 3, 4	1, 2	2	1	1, 2	1	No	No
Lim et al. (2005)	2, 3	2	5	1	1	1	Yes	No
Lobach et al. (2011)	2, 3, 4	1, 2	2	1	1	2	No	No
Lou et al. (2008)	2, 3, 4	2	5	1, 2	1, 3, 5	2	No	Yes
Mahachie John et al. (2011)	2, 3	2	4	2	3, 5	2	No	Yes
Maity et al. (2009)	2, 3, 4	2	2	1	1, 2	1	No	No
Manning et al. (2011)	2, 3	1, 2	2, 3, 5	1, 2	1	1	No	No
Mi et al. (2011)	2, 4	2	5	1, 2	2	1	No	No
Moerkerke et al. (2010)	2, 3	2	5	1	1	1	Yes	No
Mukherjee et al. (2007)	2, 3, 4	2	2	1	1, 2	1	Yes	No
Mukherjee and Chatterjee (2008)	2,3, 4	1, 2	2	1	1, 2	1	Yes	No
Mukherjee et al. (2010)	2, 3, 4	2	2	1	1, 2	1	No	No
Paré et al. (2007)	2, 3, 4	1, 2	1, 3	2	1	2	No	Yes
Ritchie et al. (2007)	2, 3	2	2	1	1, 3	2	No	Yes
Schaid (1999)	1, 2	2	2, 5	1	1, 3	1	Yes	No

Table 1 continued

Manuscript	Manuscript type[a]	Marker design[b]	Subject design[c]	Outcome type[d]	Aim[e]	Analysis approach[f]	Population stratification addressed	Applicable to gene–gene interaction
Struchalin et al. (2010)	1, 2, 3	1, 2	3, 4	2	1	1	No	No
Tan et al. (2007)	2, 3	2	5	1	1, 3	1	No	Yes
Tanck et al. (2006)	2, 3, 4	2	5	2	1, 2	2	No	Yes
Tchetgen Tchetgen and Robins (2010)	2, 3	2	1	1	1	1	No	Yes
Tzeng et al. (2011)	2, 3, 4	2	4	2	1, 3, 5	2	Yes	Yes
Umbach and Weinberg (2000)	3, 5	2	5	1	1, 5	1	Yes/no	No
Van Der Sluis et al. (2008)	2, 3	2	5	2	1, 3	1	Yes	No
Tweel and Schipper (2004)	2, 3, 4	2	2	1	1, 2	1	No	Yes
Vansteelandt et al. (2008)	2, 3	2	5	1, 2	1	1	Yes	No
Wakefield et al. (2010)	2, 3, 4	1, 2	2	1	1	2	No	Yes
Wang et el. (2009)	2, 3, 4	2	2	1	1	1	No	Yes
Witte et al. (1999)	3	2	2, 5	1	1, 3	1	Yes/no	No
Wu et al. (2009)	2, 3	2	2	1	1	1	No	No
Wyszynski and Diehl (2001)	5	2	1, 5	1	5	1	No	No
Yoshida and Koike (2011)	2, 3, 4	2	2	1	1	2	No	Yes
Yu et al. (2012)	2, 3, 4	2	2	1	1, 2	2	No	Yes
Zhang et al. (2011)	2, 3, 4	1, 2	2	1, 2	1	2	No	Yes

We have categorized the methods in terms of several features related to the type of studies or data to which the methods are meant to be applied. Some features of some methods would benefit from slightly different categorizations; we chose these as they allow the vast majority of methods to be described using similar terms. While many of the methods can or have been extended beyond what has currently been described, we limited our categorization to those situations explicitly discussed in each research paper. For example, some methods may be easily applicable to gene by gene interactions, but unless it was clear based on first principles or explicitly described in the paper, we labeled that paper "no."

Explanation coding:

[a] Overview (1), new method (2), simulated data (3), applied (4), other (5)

[b] GWAS (1), candidate gene (2), other (3)

[c] Case only (1), case/control (2), cohort (3), other-unrelateds (4), relateds (5)

[d] Binary (1), continuous (2), censored (3), multivariate (4), discrete but other than binary (5)

[e] Testing (1), modeling (2), power or sample Size (3), exposure assessment (4), other (5)

[f] Single-gene based (1), multiple gene-based (2), other (3)

Cornerstone 2: design

Similar to other epidemiologic studies, the success of *G–E* interaction studies largely depends on the selection of an optimal study design. Most common designs used for genetic association studies of main effect can be used to search for interactions. It includes family-based designs, such as nuclear families (parents and offspring) and sib designs (case and siblings), as well as common population-based designs, such as prospective cohorts and case–control data. Particular *G–E* interaction designs such as case-only designs have obtained increased popularity due to their properties and/or easy adoption. Randomized clinical trials are being curtailed to address the pharmacogenetic aspects of *G–E* interactions. However, the requirement of large sample sizes to achieve reasonable statistical power in genome-wide *G–E* interaction studies has catalyzed the development of more efficient designs over the past few years (Bookman et al. 2011). In the sequel, we briefly discuss some of the most popular designs. For a detailed summary of advantages and disadvantages of some of these designs in the context of complex trait gene–environment interaction studies, we refer to Weinberg and Umbach (2000), Dempfle et al. (2008) and Thomas (2010a).

Family-based designs can be of great interest for GWEI studies, since they usually require weaker assumptions on distributions of genetic and environmental factors than population-based designs (Liu et al. 2004). They can be more efficient when rare mutations are involved and can be robust against population stratification, although as noted earlier they still may be subject to bias in the later situation (Shi et al. 2011). Statistical tests built for family-based design are usually more robust than those built for the analysis of unrelated individuals. For example, Moerkerke et al. (2010) extended FBAT-I and established a test that is doubly robust. The approach is valid if either the model for the main genetic effect holds or if the model for the expected environmental exposure holds, but not necessarily both. Vansteelandt et al. (2008) used causal inference methodology to establish a family-based test for *G*–*E* interaction that is robust against unmeasured confounding due to population stratification and Fardo et al. (2012) extended that methodology to test for *G*–*E* interaction in family based studies with phenotypically ascertained samples.

Bias and efficiency of several family designs (e.g., using parents, siblings, cousins or "pseudo-sibs") have been studied under a range of situations by many authors (Chatterjee et al. 2005; Cordell 2009; Schaid 1999; Whittemore 2007; Witte et al. 1999). However, there is no single design that fits all purposes or is optimal for all scenario's, since utility and performance depend on disease prevalence, frequency of risk allele and risk exposure, underlying genetic model and modes of interactions, and on the goal of the study. For example, Chatterjee et al. (2005) showed some efficiency advantage of case-sibling designs compared with case-parent designs in a variety of settings. But the latter remains of interest for the estimation of the genetic association parameter (i.e., the odds ratio associated with the gene variant among subjects with environmental exposure).

Despite the advantages of family-based design, population-based design has been often preferred for genetic association studies. Ascertainment of non-relatives is logistically more convenient and potential population stratification can easily be estimated and controlled for in population-based data using genotype data from markers that are unlinked to the loci under study. Among possible population-based designs, cohort studies have long been recommended for *G*–*E* interaction studies (Clayton and McKeigue 2001). However, these remain extremely expensive and time consuming. Moreover, cohort studies are of limited use for the investigation of very rare diseases, which may require unrealizable large sample sizes. Because of this drawback, the standard case–control design (either nested in a cohort design or derived from a retrospective study) rose as the gold standard for association studies of genetic main effects (Clayton and McKeigue 2001) and is widely used in gene–environment interaction studies. Case–control designs are also often preferred to partial-collection designs (e.g., case-only, case-parents), since they might offer a better compromise between cost and efficiency (Liu et al. 2004). Statistical tests that are built within this framework are robust to a range of assumptions, such as *G*–*E* independence (although see Lindstrom et al. 2009). They generally allow unbiased estimation of all parameters that are of interest in the *G*–*E* study, although dealing with bias due to exposure misclassification remains challenging [see works from Garcia-Closas et al. for examples of impact on multiplicative interactions (Garcia-Closas et al. 1998) and impact on additive interactions (Garcia-Closas et al. 1999)].

The case-only design is probably the most discussed alternative to case-control data. It has been proposed as a less expensive design when the goal is to assess interaction effects only (Piegorsch et al. 1994; Umbach and Weinberg 1997). It relies on the assumption of independence between the genetic and environmental factor in the population. When this assumption is valid, departures from a multiplicative relative risk model can be evaluated by testing the association between *G* and *E* in cases only. This test (as well as other approaches that rely on *G*–*E* independence) has repeatedly been shown to be more efficient than other approaches. The flip side is that when the assumption does not hold, statistical tests based on cases only give rise to inflated type I error rates. Whether the aforementioned independence assumption is a reasonable one in GWEI settings is debatable. Artificial *G*–*E* dependencies can be created in multiple situations. Population stratification, for example, can create correlation between genotypes and environmental exposures in the study population (Chatterjee et al. 2005; Umbach and Weinberg 1997). Elbaz and Alperovitch (2002) have also shown that substantial correlation may appear between genetic risk factors and risk exposure of late-onset diseases in the presence of competing risks and interaction effects. Although bias in case-only designs is likely to be uncommon in practice (Dennis et al. 2011; Liu et al. 2004), using this particular design remains controversial (Albert et al. 2001). Moreover, several studies have shown that interactions opposite to the main genetic effect might not be captured within case-only data (Liu et al. 2004; Mukherjee et al. 2011).

Apart from the somewhat more traditional designs from the previous paragraphs, a range of alternative ascertainment schemes have been proposed in the literature, all with the aim to identify gene–environment interactions. Some of these designs include both related and unrelated controls (Andrieu and Goldstein 2004; Chen et al. 2009b) to increase power while others have addressed specific gene–environment interaction patterns. For example, Chen et al. (2009b) proposed a two-stage study design where a case-only study is performed at the first stage, and a case-parent/case-sibling study is performed at the second stage on a random subsample of the first-stage case sample as well as their parents/unaffected siblings. Whittemore (2007), on the other hand, discussed potential designs in studies that attempt to assess associations between lifestyle or environmental exposures and disease risk in carriers of rare mutations. Andrieu et al. (2001) also addressed the issue of rare risk factors, considering either rare mutations or rare environmental exposure. They proposed the counter-matching design which consists in increasing the number of subjects with the rare factor without increasing the number of measurements that must be performed.

Cornerstone 3: methodology

In the context of GWEI analyses, several analytical routes can be followed (Fig. 2). Some of these roads to travel by are more "natural" with specific study designs (Table 1).

Parametric and semi-parametric approaches

Many researchers have built upon the comforting regression framework in developing customized approaches to detect *G–E* interactions, including ordinary regression, penalized regression (Park and Hastie 2008), and logic regression (Schwender and Ruczinski 2010). In general, the joint effect of a genetic variant *G* and a given exposure *E* on a phenotype *Y* is often defined with the simple model:

$$g(E(Y)) = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} G \times E + \beta_{\mathbf{Z}} \mathbf{Z} \qquad (1)$$

where *G* is the number of allele (coded 0, 1, 2), *E* is continuous or categorical, **Z** represents a set of covariates one may adjust for, β is the linear effect of each component and $g()$ the link function is the logit for dichotomous *Y* and the identity for quantitative *Y*. This model is a simplification, in that it ignores possible dominance effects. Still, just as the additive model has good power over a wide range of possible dominance models and has become the primary test statistic used in most GWAS (Lettre et al. 2007), the additive main and interaction effects will be detectably non-zero for a wide range of true dominance models, and the proportion of variance explained by the missing dominance effects will be quite small for most models.

Simplification is common in classical frequentist approaches, where adding degree of freedom can reduce statistical power. Or to quote the parallel from Kooperberg and Leblanc (2008) with a cake: "if we want to divide the power over all possible interactions, nobody will get more than a crumb, and no-one will taste how good the cake is; we are better off dividing the cake among those people we believe to enjoy it." For example, a saturated linear model for a trichotomous *E* will have nine degree of freedom (*df*) compared with four *df* for the Eq. (1). In fact the same strategy has been used in most GWAS of marginal effect for the same reason.

It is important to note that even a simple model as Eq. (1) may encounter statistical issues. Especially, recent works from Tchetgen Tchetgen and Kraft (2011) have shown that when the main effect of continuous *E*, β_E in Eq. (1), is mis-specified, the likelihood ratio test, score test, and Wald test statistics of the main effect of *G* and the interaction effect can have incorrect type-1 error rates. This issue, which has been shown to be due to underestimation of the variance of β_{GE}, can be solved using different techniques (Cornelis et al. 2011): (a) using a more flexible model for the environmental main effect (e.g., adding quadratic and cubic term for the exposure); (b) using a robust ''sandwich'' estimator of the variance, and (c) modeling a continuous exposure by using general categorical variables.

A Bayesian framework gives the opportunity to make a step further in modeling the complexity of interaction effects. It provides a rational and quantitative way to consider a range of hypothesis in a single analysis. For example, Bayesian methods can be used to consider simultaneously multiple genetic models, some of them including diverse interaction effects, and to evaluate the posterior probability of each of these models [e.g., Crainiceanu et al. (2009) and Zhang and Liu (2007)]. They also allow for multiple assumptions, which can be used to build composite estimators. If one wants to quantify the relevance of the *G–E* independence assumption (discussed in further sections), they offer solutions to trade off between bias and efficiency in a data adaptive way (Li and Conti 2009; Mukherjee et al. 2010). Finally, they allow incorporating biological information and knowledge accumulated in previous association studies, so that interaction effects can be weighted by their plausibility. However, despite their potential advantages, Bayesian approaches have been only sparsely used in genetic association studies and their advantages and limits from a modeling point of view need to be studied further. In particular, many hypothesized models are likely to be roughly equally consistent with the observed data for realistic sample sizes, making it difficult to infer which model provides the best fit: the cake will be split among so many people that nobody will get more than a crumb.

Screening for variants involved in interaction when interacting factor are unknown Most genetic variants having effect through interactions with other risk factors are also likely to display marginal linear effect. For example, using random parameters for model (1) to simulate data—specifically, generating main effects and interaction effects independently of each other—will produce genetic variants with marginal effect almost 100 % of time. This suggests one can simply test for marginal effect with power being almost only related to sample size, unless (as discussed below) the state of nature is such that most true models include interaction effects, but these are offset by the main effects so that the marginal genetic effects are quite small. This is especially useful if potential interacting factor are unmeasured or when interaction effects are expected to be difficult to assess.

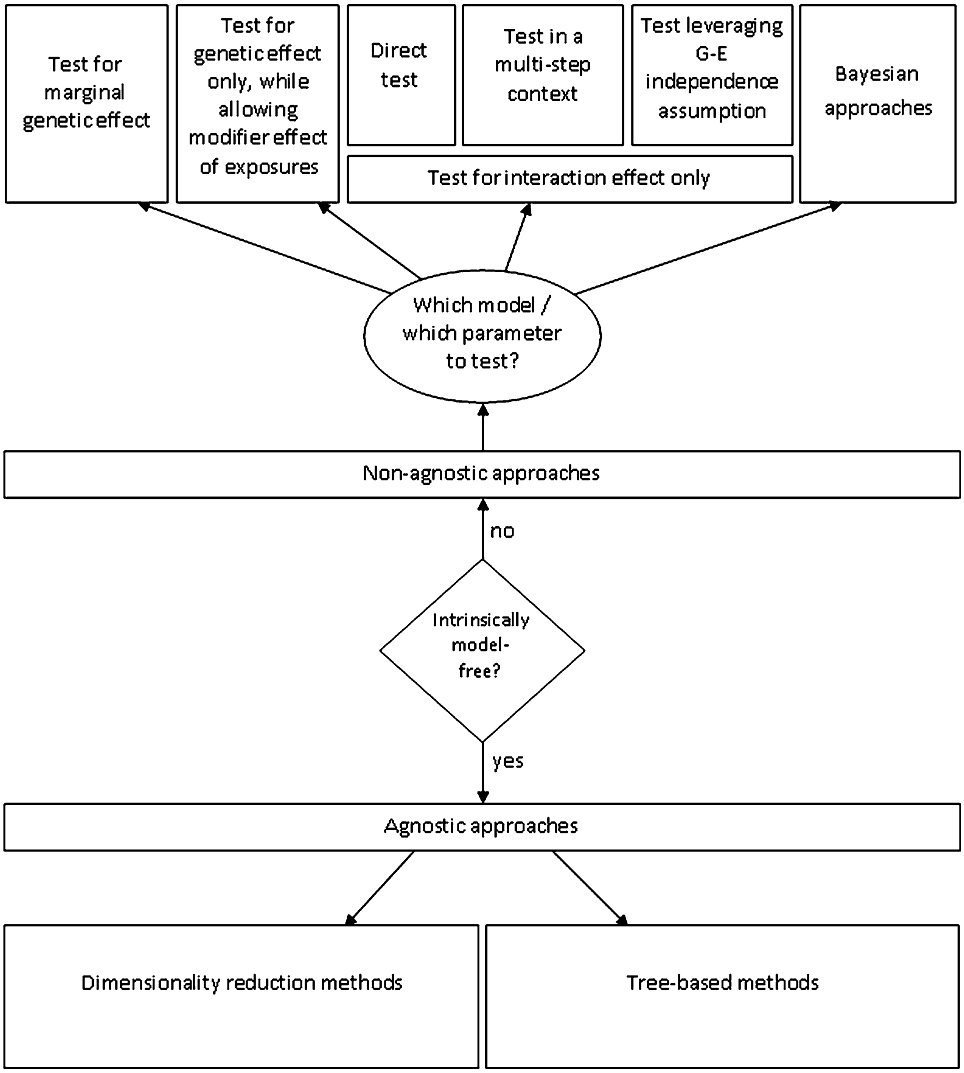
Test for marginal genetic effect
Test for genetic effect only, while allowing modifier effect of exposures
Direct test
Test in a multi-step context
Test leveraging G-E independence assumption
Bayesian approaches
Test for interaction effect only
Which model / which parameter to test?
Non-agnostic approaches
no
Intrinsically model-free?
yes
Agnostic approaches
Dimensionality reduction methods
Tree-based methods

Fig. 2 Possible strategies for GWEI depending on aim

Interaction models with small or no marginal genetic effects are theoretically possible (Culverhouse et al. 2002; Song et al. 2010). If such interactions are common, then this will have significant consequences for how we go about searching for the genetic basis of complex phenotypes and will obviously limit the interest of screening for marginal effect. However, such models have not yet been observed and confirmed in real data. This has led some to suggest that increasing sample size and testing for the marginal linear effect in agnostic GWAS scans might be the most powerful approach in most cases, while using more complex models might have only limited advantages (Clayton and McKeigue 2001; Hirschhorn and Daly 2005; Wang et al. 2005). The large success of this strategy in detecting genetic variants in GWAS has provided arguments in this direction, but the small amount of heritability explained by the "GWAS variants" is a potential rebuttal to the efficiency of this strategy.

When searching for quantitative trait loci (QTLs) an alternative for screening for the presence of interactions without using potential interacting factors is to test for homogeneity of variances across genotypic classes (Paré et al. 2010). The rationale is that, if the magnitude and the direction of the effect of a QTL differ depending on other genetic or non-genetic factors, the variability of the phenotypic outcome among individuals carrying the risk allele is likely to be larger than among the non-carrier. Hence, under the assumption that the main effect of the QTL affect neither the within-genotype variance nor the between-genotype variance, testing for heteroscedasticity will test for the presence of potential interactions. Note that heterogeneity of variances may be explained not only by the presence of interactions, but also by other biological mechanisms or other association patterns such as linkage disequilibrium with variants with large effect size (Takeuchi et al. 2011). Simulation studies have shown that the power of the test, which depend on the main effect of the unknown interacting factors (having an optimal power for specific magnitude of main effect of E), was limited when applying genome-wide significance threshold (Paré et al. 2010; Struchalin et al. 2010). Despite this limited power, testing for homogeneity of variance remains of great interest for the identification of gene–environment interactions. Because the test is potentially sensitive to a broader range of interaction effects than the test of marginal effect (such as effects in opposite direction), it can be used for example in two-step approaches to screen for candidate variants that will be tested further for gene–environment interactions. The potential of this approach has been recently demonstrated in a genome-wide association study of C-reactin and soluble ICAM-1 conducted in the Women's Genome Health Study (Paré et al. 2010). Interestingly one of the identified $G \times E$ interactions was replicated in an independent study (Dehghan et al. 2011).

Leveraging interaction effect to improve detection of marginal effect When a locus is expected to have residual marginal effects conditional on others factors tested for interaction, an efficient strategy is to use composite null hypothesis where both main effect and interaction effects are tested jointly (Kraft et al. 2007): explicitly, testing the null hypothesis that the genetic variant has no effect on any strata or based on Eq. (1) H_0: $\beta_G = 0$ or $\beta_{GE} = 0$. This can be done using a multivariate Wald test or a likelihood ratio test comparing a model including effect of E and $\mathbf{Z}$ only versus a model including effects of G, GE, E, and $\mathbf{Z}$. A simple alternative when exposure is binary or categorical is to test for marginal genetic effect in strata defined by exposure E. The joint test can then be computed as the sum of Chi-squared for association derived from each stratum. Since the samples are independents, the sum follows a Chi-square with the degree being equal to the strata for E.

For case–control studies, the test for such joint effect can be performed using standard logistic regression, the more powerful retrospective likelihood approach (Chatterjee and Carroll 2005; Cornelis et al. 2011) can exploit an underlying gene–environment independence assumption or using the empirical Bayes approach (Chen et al. 2009a; Mukherjee and Chatterjee 2008) that can data-adaptively relax the independence assumption. An extension from the family-based test for the joint test of gene main effect and G–E interaction (FBAT-J) has been recently proposed for dichotomous traits in trios and sibships (Hoffmann et al. 2009). The test assumes the genotype and the environment are independent conditionals on the parental mating type. If the assumption does not hold, the test will have an inflated type I error rate (Weinberg and Umbach 2000).

By allowing for heterogeneous genetic effect among genetic or environmental strata one can maximize the statistical power to detect the locus while minimizing the loss of power when genetic effect is homogeneous. Simulation studies have shown that a joint test for a main genetic effect and interaction effect is likely to have higher statistical power than the marginal test or the standard one degree of freedom test in presence of moderate interaction effect or when interaction effect is in opposite direction to the main effect (Kraft et al. 2007). Conversely, in the presence of a small interaction effect, the marginal test may conserve the highest power.

Methods for meta-analysis of multiple parameters have been recently described so that estimates of effects from the joint test can also be combined across independent sample. In particular, Manning et al. (2011) have described a general approach, while Aschard et al. (2011) have extended the aforementioned principle of analyzing sample stratified by environmental factors. The first approach should be used when analyzing quantitative exposures and in situations where the samples within each cohort have to be analyzed as a whole (e.g., in family data where one has to account for correlation among individuals). The second approach essentially offers practical advantage and it can be more flexible in situations where environmental categories may differ among the cohort analyzed. The first genome-wide application of the joint test has been published recently by Hamza et al. (2011). They identified a new genetic variant associated with Parkinson's disease and replicated the signal in independent samples.

As any test modeling interaction effect per se, the joint test is limited by the multiple testing issues in large-scale data. Hence, it is only applicable in situations where there is a measured factor that might interact with the tested locus. Nevertheless, some have shown that the joint test can be built in framework where multiple potential effect modifiers can be considered for a single locus. Strategy for testing can then be defined by averaging the effect of a given locus over other factors (Ferreira et al. 2007) or by testing the maximum joint test over a range of possible model (Chapman and Clayton 2007). It has been also suggested that degree-of-freedom for such joint tests can be reduced using Tukey style one-degree-of-freedom model for interaction between groups of related genetic or/and environmental variables (Chapman and Clayton 2007; Chatterjee et al. 2006; Ciampa et al. 2011).

Testing for interaction per se Besides TDT-like extension for G–E interaction as FBAT-I and its extension (Hoffmann et al. 2009; Lake and Laird 2004; Moerkerke et al. 2010) that are applicable to nuclear families data only, the traditional test for interaction consists in evaluating the term β_{GE} from Eq. (1). This test is relatively robust compared with many other approaches, although as described previously, misspecification of the main effect of a continuous E may increase type I error rate. The main concern when applying this simple test in genome-wide data is its limited power (see "Power and sample size"). Two types of strategies have been discussed to increase detection: (a) to use multi-stage approaches to reduce multiple testing burden; and (b) to leverage additional assumption on the data analyzed to improve efficiency.

Since the seminal paper from Marchini et al. (2005), multi-stage approaches using sequential test are considered as realistic approaches in GWAS. Even if not demonstrated, their work suggests that such strategy may improve the power of identifying interaction effects in GWAS. Since then diverse analysis strategies have been proposed, most of them focusing on the gene–gene interaction, which face a strong multiple testing issues in GWAS. However, these approaches can also be applied in the context of G–E. Examples include screening on genetic marginal effects (Kooperberg and Leblanc 2008; Macgregor and Khan 2006), or screening on a test that models the G–E association induced by an interaction in the combined case–control sample (Murcray et al. 2009). Simulation studies suggest that such approaches can be more powerful than traditional single-stage approach in which a huge penalty needs to be paid for multiple testing. Using a two-step strategy allows for less stringent thresholds of significance in the second step, since genetic markers have been prioritized in step one for their likely involvement in G–E interactions. While these methods became popular, questions have risen on how power and type 1 error are influenced by the correction among the two steps. While the two stages have been shown to be virtually independent in simulation study when screening on marginal effect (Kooperberg and Leblanc 2008; Marchini et al. 2005), recent work from Dai et al. (2010) provides proof of asymptotic independence of marginal association statistics and interaction statistics in linear regression, logistic regression, and Cox proportional hazard models when analyzing rare disease. Hence, in many situations the family-wise type I error rate might be controlled using classical Bonferroni correction for number of interaction tested at the second step only or by using permutation when markers considered at the second step are correlated.

Making assumption about the data analyzed to increase power of statistical test is a common principle. For binary trait such as disease status, the most popular one is the G–E independence assumption that allows testing for interaction in case-only data by testing for association between G and E among the cases using

$$\mathrm{logit}(G) = \gamma_0 + \gamma_E E + \gamma_{\mathbf{Z}} \mathbf{Z}. \quad (2)$$

Under the assumption of G–E independence in the whole population or G–E independence in controls for rare disease, testing for H_0: $\gamma_E = 0$ is equivalent to testing for H_0: $\beta_{GE} = 0$ from Eq. (1). When the assumption holds this method has the maximum power compared with most other approaches that leverage the G–E independence, except in the situation where the main effect of G or E is in opposite direction to the interaction effect (Mukherjee et al. 2011; Murcray et al. 2011). However, it has also disadvantages: the main effect of G and E cannot be estimated and the type I error can be highly inflated when the assumption does not hold.

A range of other approaches have been proposed to leverage this assumption while providing a trade-off between increased power and controlled type I error rate (Chatterjee et al. 2005; Chen et al. 2009a; Cheng 2006; Mukherjee and Chatterjee 2008; Mukherjee et al. 2007). For example, when data on both cases and controls are available in a study, then one can be much more flexible than case-only analysis in studies of gene–environment interaction irrespective of whether the independence assumption is valid or not. One can use a retrospective likelihood approach (Chatterjee et al. 2005) under the gene–environment independence assumption to obtain very efficient estimate all of the parameters of a general logistic regression model. On the other hand, if violation of the gene–environment independence assumption is suspected, one can perform data-adaptive methods such as an empirical Bayes technique (Chen et al. 2009a; Mukherjee and Chatterjee 2008), which can be robust to violation of the independence assumption and yet can be more powerful than traditional case–control analysis when the independence assumption is valid. Other alternatives to the case-only test include multi-step approaches in a single sample (Gauderman et al. 2010; Murcray et al. 2009), multi-sample design (Chen et al. 2009b), and approaches that use Bayesian framework (Li and Conti 2009; Mukherjee et al. 2010). One should note that, based on recent reports, differences in performances between these methods only exist at the margin and they always depend on the type of model simulated (see Mukherjee and Chatterjee (2008) for a detailed comparison of several of these methods).

Exploratory or agnostic approaches

Traditional statistical methods such as multivariable linear or logistic regression are ill equipped to incorporate all possible pairwise interactions among a large number of markers and exposures, let alone higher-order interactions. However, for complex diseases or traits the influence of non-linear or higher-order gene–gene and G–E interactions may be appreciable. Therefore, researchers are faced with difficult decisions to make their analysis practically feasible within computational and modeling restrictions (Maenner et al. 2009). The common alternative is to move away from the classical hypothesis testing framework and estimation of statistical significance level, and to use "model free" approaches or to adopt an agnostic approach to identify gene–environment interactions. Different analysis approaches from machine learning or data mining are needed to manage the high dimensionality of genome-wide analysis studies and large-scale data collections.

Interdisciplinary collaborations have led to the adoption of approaches from one community to another, especially in the field of gene–gene interactions. These include data segmentation methods (Tryon 1939), tree-based methods (Breiman et al. 1984), pattern recognition methods (Ripley 1996), and (non-)linear dimension reduction methods (Fodor 2002). A list of examples of these in the context of gene–gene interactions is given in Van Steen (2012). Unfortunately, the adoption of these methods in genome-wide based *G–E* interaction detection is not as "frequent" as it is for genome-wide epistasis studies. In the following, we elaborate on two techniques that deserve more attention in the context of GWEI studies: tree-based and multifactor dimensionality reduction (MDR) derived techniques.

Because the number of possible genetic model can be quite large, exploratory methods are often built on a trade-off and assume or favor some specific interaction models. Recursive partitioning approaches, such as random forests (Breiman et al. 1984; Schwarz et al. 2010)—a flexible and efficient data mining method based on regression or classification trees—also face such issues. Random forests do not model interaction variables per se but they allow for interactions (or complex non-linear relationships) in the sense that they evaluate classification ability of particular combination of values taken by sets of predictor variables. Because of the independence assumption used during node splitting of "trees" these methods have been shown to have limited ability to detect pure interaction effects (McKinney et al. 2009). Notably, the recent SNPInterForest approach (Yoshida and Koike 2011) performed very well in successfully identifying pure epistatic interactions with high precision and was still more than capable of concurrently identifying multiple interactions under the existence of genetic heterogeneity. Hence, extensions that relax the independence assumption within a conditional inference framework (Hothorn et al. 2006) and improved procedures to extract interaction patterns from random forest (Yoshida and Koike 2011) make the random forest methodology particularly attractive for GWEI studies. Different variable importance measures have been proposed in the literature, including a joint importance measure which extends the idea of single importance to multiple importance, and can be useful especially for interactions (Bureau et al. 2005). Note that correlated predictors and varying predictor categories or measurement scales are likely to exist in *G–E* studies and that care needs to be taken in the selection of the importance criterion. For instance, Strobl et al. (2008) identified the mechanisms causing the bias for permutation importance scores and developed a conditional variable importance which reflects the true impact of each predictor variable more reliable than the original permutation variable importance measure.

As an application example, Maenner et al. (2009) analyzed coronary heart disease cases from the Framingham Heart Study by first identifying influential SNPs for age of onset of early coronary heart disease using a random forest approach. Variable importance scores from a RF analysis provide measures to determine important SNPs and environmental exposures taking into account interactions without specifying a genetic model (Lunetta et al. 2004). Second, generalized estimating equations were used to evaluate the statistical significance of main effects and interactions of previously detected SNPs and smoking status (Maenner et al. 2009) (however, note that such significance level should be taken with caution since the selection at the "mining step" potentially overfits the data). The authors used a simple solution to handle family structure within their data by considering a binary family indicator as covariate for building the random forest. Similarly, Zhai et al. (2011) performed a two-step approach with initial screening for SNPs associated with environmental measures by random forest and further analysis based on case-only logistic regression to obtain parameter estimates for the selected variables.

Tree-based methods might be a relevant alternative to logistic regression methods for identifying genes without strong marginal effects and of robustness to genetic heterogeneity where different subsets of genes can lead to a phenotype of interest (Lunetta et al. 2004). Random forests outperformed Fisher's exact test when several risk SNPs interact (Lunetta et al. 2004) and behave more robustly when a high number of unassociated noise SNPs is present (Bureau et al. 2005). Another interesting approach combining regression models and tree-based methodology is a semi-parametric regression model, named partially linear tree-based regression model (PLTR) (Chen et al. 2007). The linear regression part of their model can control efficiently for confounders and provide the possibility to correct for linear main effects of variables so that a parsimonious summary of the joint effect of genetic and environmental variables is obtained.

Also non-parametric data mining methods such as MDR (Ritchie et al. 2001) are the subjects of a trade-off. In contrast to logistic regression and random forests, MDR can be used to detect *G–E* interactions in the absence of any main effects. MDR can be applied to smaller sample sizes than logistic regression which needs enough observations to model all main and interactions effects. However, the "reduction" step consists in splitting the different combination of two variables (defined by *E* and *G*) in two groups of high risk versus low risk. This allows a range of model to be tested. But the interaction is summarized in a single binary parameter and is therefore unlikely to capture the full complexity of interactions (e.g., a gradient of effect across different combinations). Several extensions and variations of the MDR method have been proposed to address initial shortcomings of MDR (including the lack of correction for lower order effects, and the too stringent reduction into two risk groups). Model-based MDR (MB-MDR) (Calle et al. 2010) and its extension to family data, family MDR (FAM-MDR) (Cattaert et al. 2010), enable adjustments for possible confounders and the handling of various phenotypes, e.g., continuous, categorical, or censored. In particular, MB-MDR uses a reduction into a one-dimensional variable with three levels, i.e., high risk, no evidence, low risk, and potentially a continuum of risk groups (Calle et al. 2010; Cattaert et al. 2010). While comparing MB-MDR to MDR in the presence of noise, i.e., genotyping error, phenocopy and genetic heterogeneity, MB-MDR was found to have increased power in most situations, especially for genetic heterogeneity, phenocopies, and minor allele frequencies. Previous to applying the MB-MDR method, FAM-MDR uses a preparation step where familial correlation free traits are obtained as residuals from a polygenic model (hence, hereby adjusting for potential population stratification). FAM-MDR outperformed pedigree-based MDR (PGMDR) (Lou et al. 2008) in terms of handling multiple testing, empirical power, and efficient use of available information from complex and extended pedigrees (Cattaert et al. 2010) and is therefore a promising alternative to the classical MDR derivatives to explore gene–environment interactions. One disadvantage of MDR is that its computational burden increases with the number of SNPs and the order of considered interactions. A parallel algorithm of MDR and MB-MDR has been implemented by Bush et al. (2006) and Van Lishout et al. (2011), respectively. Despite these efforts, filtering methods to preselect a subset of candidate factors and stochastic search algorithms (e.g., simulated annealing and evolutionary algorithms) are needed to assist researchers in the exhaustive search for interactions in genome-wide association studies. Knowledge about the pros and cons of these filtering approaches (as applied to genome-wide epistasis settings) will be most beneficial for GWEI studies and the availability of an entire exposome.

Duell et al. (2008) compared MDR to focused interaction testing framework and logistic regression for identification of higher-order interaction effects in a case–control study using 26 polymorphisms and smoking as possible environmental risk factor. Little concordance existed between MDR and interaction testing framework with regard to the interaction factors. This finding may be caused by the different interaction modeling methodologies behind the approaches. The authors recommend using multiple approaches for data screening and analysis to detect potentially new risk factor combinations. More comparative studies are actually needed, examining differences between traditional (often regression-based) approaches with untraditional (often data-mining) methods in the context of GWEI studies. The study from Duell et al. (2008) also highlights the difficulties in computing a comprehensive significance level for exploratory methods. Overall, one should remember that there is no straightforward way to define a null hypothesis and to test it in these exploratory approaches. However, strategies to statistically evaluate the significance of models obtained through data mining procedures are now discussed in the literature (e.g., Pattin et al. 2009) and more might be developed in the future.

Out-of-the-box approaches

Information theoretic metrics allow for complex interactions between genetic variations and environmental factors without any modeling but have not yet been widely applied. Based on the total correlation information (TCI) (Chanda et al. 2007), Chanda et al. (2008) developed the phenotype-associated information (PAI), which is robust against dependencies between environmental and genetic factors. Furthermore, these authors suggest a greedy search algorithm (AMBIENCE) where potential variable combinations associated with a phenotype of interest are selected based on lower order PAI values and the interaction between the determined relevant variable subsets is re-evaluated using the more parsimonious *k*-way interaction information. This approach is particularly suitable for large-scale data sets. The method was extended to quantitative traits (Chanda et al. 2009a), when normally distributed within each strata of the gene–environmental variable combination. Wu et al. (2009) and Fan et al. (2011) used test-statistics developed from information theoretic metrics to detect *G–E* interactions associated with discrete phenotypes. While the mutual information-based test statistic of Wu et al. (2009) is applicable to two-way interactions, Fan et al. (2011) also consider higher order interactions. An extension of their computationally efficient approaches to quantitative traits and family data would increase the applicability and flexibility of information theoretic metrics further.

To prioritize genetic and environmental variables for follow-up sequencing studies, Chanda et al. (2009b) proposed to calculate the interaction index (defined as the sum of the average interaction contribution of each considered *k*th order interaction for the given variable) for each variable. Comparing their approach with the restricted partitioning method (RPM) (Culverhouse et al. 2004), Chanda et al. (2009b) found high concordance between the two methods for one-variable combinations but not for the two-variable combinations. In contrast to, for instance, MDR and RPM, the greedy search algorithm AMBIENCE (Chanda et al. 2008) allows for higher dimensional datasets but disables the detection of pure epistasis effects. An alternative approach to the search algorithm might be to use an information-theoretic metrics as objective function in a dimensionality reduction method as MDR for which variables could be pooled into high-risk and low-risk sets based on their PAI value (Fodor 2002).

Recently, rule-based classifier algorithms have been introduced in the context of genetic interaction studies, whereas they had proven their utility non-genetic datasets in the past (Tan et al. 2006). Rule-based classifiers generate classification models using a collection of "if … then …" rules. The algorithms are computationally feasible and allow the inclusion of both categorical and continuous variables. For a comparison of rule-based classifiers in the context of *G–E* interactions, we refer to Lehr et al. (2011).

Alternatively, GWEI studies may benefit from neural networks (NN) (Gunther et al. 2009) and their modifications, e.g., genetic programming neural networks (GPNN) (Ritchie et al. 2007) and grammatical evolution neural networks (GENN) (Motsinger et al. 2006).

Unlike logistic regressions, neural networks do not explicitly use interaction terms for modeling data. There is no easy way to assess whether interaction is present using a neural network, nor to derive clear interpretations of estimated weights (Gunther et al. 2009). The GPNN algorithm attempts to generate optimal neural network architecture for a given data set and—in contrast to classical NN—does not rely on the pre-specification of inputs and architecture (Ritchie et al. 2007). Although these types of approaches are often regarded as a black box, the flexibility of neural network-based approaches in model development clearly is a major advantage, especially when highly complex data structures with challenging gene–gene or *G–E* interaction structures need to be modeled.

GWEI and GWAI studies

Large-scale *G–E* interaction studies and large-scale gene–gene interaction studies, via the common genetic component they involve, share quite a number of challenges: high-dimensionality, computational capability, the absence/presence of marginal effects, the multiple testing problem, and genetic heterogeneity. These challenges and possible solutions in the context of genome-wide association gene–gene interaction (GWAI) studies have been discussed elsewhere (Van Steen 2012).

When environmental risks are investigated, usually the focus is on a single exposure or several exposures from particular category, for instance, involving air and water pollution, occupation, diet, stress and behavior, or types of infection. However, in the context of a genome-wide screen for loci involved in interactions, a marker may interact with an exposure from any category, or multiple exposures within or across categories. The effect of a marker may differ across strata defined by more than one exposure (e.g., the effect of a breast cancer marker might be different among women with a high Gail score, which summarizes several non-genetic breast-cancer risk factors, and women with a low Gail score). Along those lines, it is believed to be crucial to combine the genome with an entire "exposome" (i.e., the totality of environmental exposures from conception onwards) (Wild 2005). This idea is similar to evaluating the effects of genetic variants in a particular genetic background, as summarized by high-dimensional genetic data (Phillips 2008; Tzeng et al. 2011; Van Steen 2012). Methods for the measurement of the "exposome" are lagging far behind methods for measuring genomic variation. However, instead of characterizing the entire exposome, it should be feasible to identify at least critical components at several stages in an individual's life and consider these in the *G–E* analysis (Rappaport and Smith 2010). The Bayesian paradigm is promising in this sense, since latent variables can potentially be used to capture genetic variation and models can be developed allowing environment effects to vary across different genetic profile categories (Yu et al. 2012).

GWEI studies may benefit from the abundance of methodologies that are available in the context of large-scale genetic association or epistasis screenings (Khoury and Wacholder 2009). We believe that there are several reasons for the limited translation of GWAI to GWEI methodologies. First, genome-wide *G–E* interaction studies have only recently become possible through several organized large-scale data collections (Davis and Khoury 2007) containing both genetic and good quality environmental measurements. Still, germline variations are static and can be captured at any time point, while exposures can change over time and are not always measured at the relevant time period (measurement at baseline or at interview may not reflect the relevant windows of exposure and will not reflect lifetime exposure). Hence some GWAI methods are likely to be underpowered since they are not designed to account for such variations. Second, GWEI studies involve factors that are measured on different scales. GWAI studies usually involve one type of genetic markers that have been pre-processed and underwent high quality control procedures. The measurement type (coding) is regarded to be the same for all SNPs in the analysis. An environmental factor can be continuous, categorical, or binary, whatever reflects the true underlying nature best. Combining different measurement scales within one approach, and inclusion of factors with differential degrees of accuracy, measurement error or variability poses additional complications [e.g., in random forests approaches (Strobl et al. 2007, 2008)]. Third, for GWAI studies there is a consensus on how to deal with missing genotypes. Several procedures have been developed to "impute" missing data in this context, for instance, using HapMap reference data. Clearly, the taxonomy of Little and Rubin (1987) and bio-statistical knowledge about missing data handling in epidemiology now need to be combined with missing data handling techniques commonly adopted in statistical genetics. We refer to recent work of Lobach et al. (2011) that discusses exposure measurement error and genotype missing data in the context of a small-scaled gene–environment interaction analysis. Fourth, GWEI studies may face additional methodological challenges when the original GWAS study is based on shared publicly available controls. It has been now well established that use of shared controls, after appropriate adjust for population stratification using principal component and related methods, produces valid inference for detection of genetic main effects. For studies of gene–environment interaction, however, one needs more caution as the exposure distribution for the underlying population of the controls may be quite different from the exposure distribution for the underlying population from which the cases were drawn. Further, data on relevant environmental exposures of interest may not often be available on publicly available studies. In such situation, one can use a case-only analysis to examine multiplicative gene–environment interaction, but such inference is inherently limited as we have noted earlier. Fifth, meta-analytic approaches to boost power of GWEI studies are usually limited to parametric *G–E* detection methods that result in estimable effect sizes (Aschard et al. 2011; Manning et al. 2011). Model misspecification is one of the major concerns in meta-analysis contexts (Pereira et al. 2011; Pereira et al. 2009). General approaches are needed that require no assumption on modes of action in the meta-analytical context of GWEI studies. Finally, meta-GWEI studies will further benefit from continuing efforts to improve the accuracy of epidemiological questionnaires of medical records, occupational records, and other proxy measurements of environmental factors, as well as the development of low-cost, validated, and standardized environmental measures, (Bookman et al. 2011; Khoury and Wacholder 2009).

Future perspectives

The detection of *G–E* interactions is usually based on making inferences from statistical interactions that are observed at a population level, the most popular methodologies being based on regression paradigms. The most interesting types of *G–E* interactions are those that are coined "non-removable", in the sense that the evidence of (statistical) interaction exists when no obvious monotone transformation of the trait exists (i.e., rescaling of the trait) that removes the interaction. Uher (2008) argued that concerns about statistical models and scaling can be addressed by integration of observed and experimental data, assuming, however, that we already have identified "interesting" environmental risk factors. Most of these risk factors for common complex diseases have not yet been identified, and for those that have been identified, the mode of action is not well known. Moving from a hypothesis-driven to a hypothesis-generating viewpoint (i.e., from a limited selection of candidate environmental risk factors to an exposome) magnifies some of the issues involved in interaction detection, with agents that may be highly structured or inter-connected in epidemiological or biological networks. Fortunately, lessons can be learnt from similar settings, such as those generated by GWAI data. Several efforts are being made to tackle some of the identified hurdles in this manuscript (Engelman et al. 2009) and a steady increase in GWEI studies is observed (refer to Fig. 1). Although most of the identified interactions have not yet been confirmed, the first GWEI results suggest the importance of testing for *G–E* interactions. Adopting an interdisciplinary attitude and a systems biology view, using out-of-the-box strategies and non-linear mathematics that are less known in epidemiology (Knox 2010) may help identify interacting factors and better understand gene–environment interplay.

A *G–E* interaction effect in a population is dependent upon the distribution of genetic and environmental factors in the population of interest. Obviously, the distribution of environmental and genetic factors can be quite different between individuals and across populations. Thus, some *observed G–E* interaction effects, including those involving epigenetic phenomena, might be detected in one population but be absent in another. We wish to emphasize that in valid epidemiologic comparisons, controls should be a random sample of the population from which the cases arise. If a control were to become a case, would he or she be selected as a case in your study?

The availability of the entire sequence of the human genome offers enormous opportunities. It is now possible to obtain data on rare variants as well as common variants, for complex disease association studies. The effects on dimensionality are enormous, but Wray et al. (2011) have argued that genes identified via GWAS studies harboring common variants are likely to be good candidates for the identification of rare variants, which can then (theoretically) be investigated for their relationship with a disease trait. The role of rare variants (relative to more common variants) in complex disease etiology is still unclear. It has been proposed that multiple rare variants through LD may be responsible for some of the common variant hits from recent GWAS (so called synthetic associations); however, this has been deemed unlikely (Anderson et al. 2011). Large-scale sequencing efforts will be required to fully investigate the genetic architecture of complex disease etiology. Understanding how one or more rare variants may interact with each other and with environmental exposures will be an extremely difficult task to accomplish. Many thousands of participants will be required even to evaluate main effects of rare variants. The analysis of interactions between rare variants and environmental exposures will be very challenging for the same reasons it is difficult for common variants. Moreover, if we are willing to believe that most chronic diseases are a result of numerous subtle perturbations in exogenous and endogenous exposures and variation at the epigenomic level, then each individual may indeed have their own ‘personalized interactome’. This could have tremendous implications for the study of *G–E* and *G–G* interactions and might help to explain why even very large consortium efforts have been unsuccessful at identifying more than a minor fraction of the heritability of disease.

Acknowledgments H. Aschard and P. Kraft were supported by grant NIH/NIDDK-R21 DK084529. B. Maus and K. Van Steen acknowledge research opportunities offered by the Belgian Network BioMAGNet (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Program (Phase VI/4), initiated by the Belgian State, Science Policy Office. Their work was also supported in part by the IST Program of the European Community, under the PASCAL2 Network of Excellence (Pattern Analysis, Statistical Modeling and Computational Learning), IST-2007-216886. E.J. Duell was supported by the Spanish Ministry of Health (ISCIII RETICC RD06/0020). The scientific responsibility for this work rests with its authors.

Conflict of interest The authors declare that they have no conflict of interest.

References

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S (2001) Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154:687–693

Albrechtsen A, Castella S, Andersen G, Hansen T, Pedersen O, Nielsen R (2007) A Bayesian multilocus association method: allowing for higher-order interaction in association studies. Genetics 176:1197–1208

Amato R, Pinelli M, D’Andrea D, Miele G, Nicodemi M, Raiconi G, Cocozza S (2010) A novel approach to simulate gene-environment interactions in complex diseases. BMC Bioinformatics 11:8

Anderson CA, Soranzo N, Zeggini E, Barrett JC (2011) Synthetic associations are unlikely to account for many common disease genome-wide association signals. Plos Biol 9:e1000580

Andrieu N, Goldstein AM (2004) The case-combined-control design was efficient in detecting gene-environment interactions. J Clin Epidemiol 57:662–671

Andrieu N, Goldstein AM, Thomas DC, Langholz B (2001) Counter-matching in studies of gene-environment interaction: efficiency and feasibility. Am J Epidemiol 153:265–274

Aschard H, Hancock DB, London SJ, Kraft P (2011) Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. Hum Hered 70:292–300

Aschard H, Chen J, Cornelis M, Chibnik L, Karlson E, Kraft P (2012) Inclusion of gene–gene and gene–environment interactions unlikely to dramatically improve risk prediction for complex diseases. Am J Hum Genet 90(6):962–972

Balansky R, Ganchev G, Iltcheva M, Nikolov M, Steele VE, De Flora S (2012) Differential carcinogenicity of cigarette smoke in mice exposed either transplacentally, early in life or in adulthood. Int J Cancer 130:1001–1010

Bashir SA, Duffy SW (1995) Correction of risk estimates for measurement error in epidemiology. Methods Inf Med 34:503–510

Bhattacharjee S, Wang Z, Ciampa J, Kraft P, Chanock S, Yu K, Chatterjee N (2010) Using principal components of genetic variation for robust and powerful detection of gene–gene interactions in case-control and case-only studies. Am J Hum Genet 86:331–342

Bookman EB, McAllister K, Gillanders E, Wanke K, Balshaw D, Rutter J, Reedy J, Shaughnessy D, Agurs-Collins T, Paltoo D, Atienza A, Bierut L, Kraft P, Fallin MD, Perera F, Turkheimer E, Boardman J, Marazita ML, Rappaport SM, Boerwinkle E, Suomi SJ, Caporaso NE, Hertz-Picciotto I, Jacobson KC, Lowe WL, Goldman LR, Duggal P, Gunnar MR, Manolio TA, Green ED, Olster DH, Birnbaum LS (2011) Gene-environment interplay in common complex diseases: forging an integrative model-recommendations from an NIH workshop. Genet Epidemiol 35:217–225

Bouzigon E, Corda E, Aschard H, Dizier MH, Boland A, Bousquet J, Chateigner N, Gormand F, Just J, Le Moual N, Scheinmann P, Siroux V, Vervloet D, Zelenika D, Pin I, Kauffmann F, Lathrop M, Demenais F (2008) Effect of 17q21 variants and smoking exposure in early-onset asthma. N Engl J Med 359:1985–1994

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman & Hall/CRC, New York

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P (2005) Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 28:171–182

Bush WS, Dudek SM, Ritchie MD (2006) Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene–gene interactions. Bioinformatics 22:2173–2174

Bůžková P, Lumley T, Rice K (2011) Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. Ann Hum Genet 75(1):36–45

Calle ML, Urrea V, Malats N, Van Steen K (2010) mbmdr: an R package for exploring gene–gene interactions associated with binary or quantitative traits. Bioinformatics 26:2198–2199

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models, 2nd edn. Chapman & Hall/CRC Press, Boca Raton

Cattaert T, Urrea V, Naj AC, De Lobel L, De Wit V, Fu M, Mahachie John JM, Shen H, Calle ML, Ritchie MD, Edwards TL, Van Steen K (2010) FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals. PLoS One 5:e10304

Chanda P, Zhang A, Brazeau D, Sucheston L, Freudenheim JL, Ambrosone C, Ramanathan M (2007) Information-theoretic metrics for visualizing gene-environment interactions. Am J Hum Genet 81:939–963

Chanda P, Sucheston L, Zhang A, Brazeau D, Freudenheim JL, Ambrosone C, Ramanathan M (2008) AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. Genetics 180:1191–1210

Chanda P, Sucheston L, Liu S, Zhang A, Ramanathan M (2009a) Information-theoretic gene–gene and gene-environment interaction analysis of quantitative traits. BMC Genomics 10:509

Chanda P, Sucheston L, Zhang A, Ramanathan M (2009b) The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors. Eur J Hum Genet 17:1274–1286

Chapman J, Clayton D (2007) Detecting association using epistatic information. Genet Epidemiol 31:894–909

Chatterjee N, Carroll RJ (2005) Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92:399–418

Chatterjee N, Kalaylioglu Z, Carroll RJ (2005) Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. Genet Epidemiol 28:138–156

Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S (2006) Powerful multilocus tests of genetic association in the presence of gene–gene and gene-environment interactions. Am J Hum Genet 79:1002–1016

Chen J, Yu K, Hsing A, Therneau TM (2007) A partially linear tree-based regression model for assessing complex joint gene–gene and gene-environment effects. Genet Epidemiol 31:238–251

Chen YH, Chatterjee N, Carroll RJ (2008) Retrospective analysis of haplotype-based case control studies under a flexible model for gene environment association. Biostatistics 9:81–99

Chen YH, Chatterjee N, Carroll RJ (2009a) Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. J Am Stat Assoc 104:220–233

Chen YH, Lin HW, Liu HM (2009b) Two-stage Analysis for gene-environment interaction utilizing both case-only and family-based analysis. Genet Epidemiol 33:95–104

Cheng KF (2006) A maximum likelihood method for studying gene-environment interactions under conditional independence of genotype and exposure. Stat Med 25:3093–3109

Ciampa J, Yeager M, Jacobs K, Thun MJ, Gapstur S, Albanes D, Virtamo J, Weinstein SJ, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Hunter D, Hoover R, Thomas G, Chanock S, Holmes C, Chatterjee N (2011) Application of a novel score test for genetic association incorporating gene–gene interaction suggests functionality for prostate cancer susceptibility regions. Hum Hered 72:182–193

Clayton DG (2009) Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet 5:e1000540

Clayton D, McKeigue PM (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358:1356–1360

Colilla S, Nicolae D, Pluzhnikov A, Blumenthal MN, Beaty TH, Bleecker ER, Lange EM, Rich SS, Meyers DA, Ober C, Cox NJ, Asthm CSG (2003) Evidence for gene-environment interactions in a linkage study of asthma and smoking exposure. J Allergy Clin Immunol 111:840–846

Cordell HJ (2009) Estimation and testing of gene-environment interactions in family-based association studies. Genomics 93:5–9

Cordell HJ, Barratt BJ, Clayton DG (2004) Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene–gene and gene-environment interactions, and parent-of-origin effects. Genet Epidemiol 26:167–185

Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, Bennett SN, Bierut LJ, Boerwinkle E, Doheny KF, Feenstra B, Feingold E, Fornage M, Haiman CA, Harris EL, Hayes MG, Heit JA, Hu FB, Kang JH, Laurie CC, Ling H, Manolio TA, Marazita ML, Mathias RA, Mirel DB, Paschall J, Pasquale LR, Pugh EW, Rice JP, Udren J, van Dam RM, Wang X, Wiggs JL, Williams K, Yu K (2010) The Gene, Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet Epidemiol 34:364–372

Cornelis MC, Tchetgen Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P (2011) Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. Am J Epidemiol 175:191–202

Crainiceanu A, Liang KY, Crainiceanu CM (2009) Bootstrap Bayesian analysis with applications to gene-environment interaction. In: 24th International Symposium on Computer and Information Sciences, pp 649–654

Culverhouse R, Suarez BK, Lin J, Reich T (2002) A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 70:461–471

Culverhouse R, Klein T, Shannon W (2004) Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 27:141–152

Dai JY, Kooperberg C, LeBlanc M, Prentice RL (2010) On two-stage hypothesis testing procedures via asymptotically independent statistics. UW Biostatistics Working Paper Series. Working Paper 367

Davis RL, Khoury MJ (2007) The emergence of biobanks: practical design considerations for large population-based studies of gene-environment interactions. Community Genet 10:181–185

Dehghan A, Dupuis J, Barbalic M, Bis JC, Eiriksdottir G, Lu C, Pellikka N, Wallaschofski H, Kettunen J, Henneman P, Baumert J, Strachan DP, Fuchsberger C, Vitart V, Wilson JF, Pare G, Naitza S, Rudock ME, Surakka I, de Geus EJ, Alizadeh BZ, Guralnik J, Shuldiner A, Tanaka T, Zee RY, Schnabel RB, Nambi V, Kavousi M, Ripatti S, Nauck M, Smith NL, Smith AV, Sundvall J, Scheet P, Liu Y, Ruokonen A, Rose LM, Larson MG, Hoogeveen RC, Freimer NB, Teumer A, Tracy RP, Launer LJ, Buring JE, Yamamoto JF, Folsom AR, Sijbrands EJ, Pankow J, Elliott P, Keaney JF, Sun W, Sarin AP, Fontes JD, Badola S, Astor BC, Hofman A, Pouta A, Werdan K, Greiser KH, Kuss O, Meyer zu Schwabedissen HE, Thiery J, Jamshidi Y, Nolte IM, Soranzo N, Spector TD, Volzke H, Parker AN, Aspelund T, Bates D, Young L, Tsui K, Siscovick DS, Guo X, Rotter JI, Uda M, Schlessinger D, Rudan I, Hicks AA, Penninx BW, Thorand B, Gieger C, Coresh J, Willemsen G, Harris TB, Uitterlinden AG, Jarvelin MR, Rice K, Radke D, Salomaa V, Willems van Dijk K, Boerwinkle E, Vasan RS, Ferrucci L, Gibson QD, Bandinelli S, Snieder H, Boomsma DI, Xiao X, Campbell H et al (2011) Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation 123:731–738

Demissie S, Cupples LA (2011) Bias due to two-stage residual-outcome regression analysis in genetic association studies. Genet Epidemiol 35:592–596

Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, Schafer H (2008) Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet 16:1164–1172

Dennis J, Hawken S, Krewski D, Birkett N, Gheorghe M, Frei J, McKeown-Eyssen G, Little J (2011) Bias in the case-only design applied to studies of gene-environment and gene–gene interaction: a systematic review and meta-analysis. Int J Epidemiol 40:1329–1341

Dizier MH, Selinger-Leneman H, Genin E (2003) Testing linkage and gene × environment interaction: comparison of different affected sib-pair methods. Genet Epidemiol 25:73–79

Doherty SP, Grabowski J, Hoffman C, Ng SP, Zelikoff JT (2009) Early life insult from cigarette smoke may be predictive of chronic diseases later in life. Biomarkers 14(Suppl 1):97–101

Duell EJ, Bracci PM, Moore JH, Burk RD, Kelsey KT, Holly EA (2008) Detecting pathway-based gene–gene and gene-environment interactions in pancreatic cancer. Cancer Epidemiol Biomarkers Prev 17:1470–1479

Dunn EC, Uddin M, Subramanian SV, Smoller JW, Galea S, Koenen KC (2011) Research review: gene-environment interaction research in youth depression—a systematic review with recommendations for future research. J Child Psychol Psychiatry 52:1223–1238

Efird JT (2005) Method for indirectly estimating gene-environment effect modification and power given only genotype frequency and odds ratio of environmental exposure. Eur J Epidemiol 20:389–393

Ege MJ, Strachan DP, Cookson WO, Moffatt MF, Gut I, Lathrop M, Kabesch M, Genuneit J, Buchele G, Sozanska B, Boznanski A, Cullinan P, Horak E, Bieli C, Braun-Fahrlander C, Heederik D, von Mutius E (2011) Gene-environment interaction for childhood asthma and exposure to farming in Central Europe. J Allergy Clin Immunol 127:138–144, 144.e1–144.e4

Elbaz A, Alperovitch A (2002) Bias in association studies resulting from gene-environment interactions and competing risks. Am J Epidemiol 155:265–272

Engelman CD, Baurley JW, Chiu YF, Joubert BR, Lewinger JP, Maenner MJ, Murcray CE, Shi G, Gauderman WJ (2009) Detecting gene-environment interactions in genome-wide association data. Genet Epidemiol 33:S68–S73

Fan R, Zhong M, Wang S, Zhang Y, Andrew A, Karagas M, Chen H, Amos CI, Xiong M, Moore JH (2011) Entropy-based information gain approaches to detect and to characterize gene–gene and gene-environment interactions/correlations of complex diseases. Genet Epidemiol 35:706–721

Fardo DW, Liu J, Demeo DL, Silverman EK, Vansteelandt S (2012) Gene-environment interaction testing in family-based association studies with phenotypically ascertained samples: a causal inference approach. Biostatistics 13:468–481

Ferreira T, Donnelly P, Marchini J (2007) Powerful Bayesian gene–gene interaction analysis. Am J Hum Genet S81:32

Fodor I (2002) A survey of dimension reduction techniques. LLNL technical report

Franks PW (2011) Gene × environment interactions in type 2 diabetes. Curr Diab Rep 11:552–561

Garcia-Closas M, Thompson WD, Robins JM (1998) Differential misclassification and the assessment of gene-environment interactions in case-control studies. Am J Epidemiol 147:426–433

Garcia-Closas M, Rothman N, Lubin J (1999) Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. Cancer Epidemiol Biomarkers Prev 8:1043–1050

Gauderman WJ, Faucett CL (1997) Detection of gene-environment interactions in joint segregation and linkage analysis. Am J Hum Genet 61:1189–1199

Gauderman WJ, Thomas DC (2001) The role of interacting determinants in the localization of genes. Adv Genet 42:393–412

Gauderman WJ, Thomas DC, Murcray CE, Conti D, Li D, Lewinger JP (2010) Efficient genome-wide association testing of gene-environment interaction in case-parent trios. Am J Epidemiol 172:116–122

Geneletti S, Gallo V, Porta M, Khoury MJ, Vineis P (2011) Assessing causal relationships in genomics: from Bradford-Hill criteria to complex gene-environment interactions and directed acyclic graphs. Emerg Themes Epidemiol 8:5

Gibson G (2010) Hints of hidden heritability in GWAS. Nat Genet 42:558–560

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

Gu CC, Yang WW, Kraja AT, de Las Fuentes L, Davila-Roman VG (2009) Genetic association analysis of coronary heart disease by profiling gene-environment interaction based on latent components in longitudinal endophenotypes. BMC Proc 3(Suppl 7):S86

Gunther F, Wawro N, Bammann K (2009) Neural networks for modeling gene–gene interactions in association studies. BMC Genet 10:87

Hamza TH, Chen H, Hill-Burns EM, Rhodes SL, Montimurro J, Kay DM, Tenesa A, Kusel VI, Sheehan P, Eaaswarkhanth M, Yearout D, Samii A, Roberts JW, Agarwal P, Bordelon Y, Park Y, Wang L, Gao J, Vance JM, Kendler KS, Bacanu SA, Scott WK, Ritz B, Nutt J, Factor SA, Zabetian CP, Payami H (2011) Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. PLoS Genet 7:e1002237

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Hoffmann TJ, Lange C, Vansteelandt S, Laird NM (2009) Gene-environment interaction tests for dichotomous traits in trios and sibships. Genet Epidemiol 33:691–699

Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 15:651–674

Hunter DJ (2005) Gene-environment interactions in human diseases. Nat Rev Genet 6:287–298

Karageorgi S, Prescott J, Wong JY, Lee IM, Buring JE, De Vivo I (2011) GSTM1 and GSTT1 copy number variation in population-based studies of endometrial cancer risk. Cancer Epidemiol Biomarkers Prev 20:1447–1452

Kazma R, Babron MC, Genin E (2011) Genetic association and gene-environment interaction: a new method for overcoming the lack of exposure information in controls. Am J Epidemiol 173:225–235

Khoury MJ, Wacholder S (2009) Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies—challenges and opportunities. Am J Epidemiol 169:227–230 (discussion 234–235)

Knox SS (2010) From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. Cancer Cell Int 10:11

Kooperberg C, Leblanc M (2008) Increasing the power of identifying gene × gene interactions in genome-wide association studies. Genet Epidemiol 32:255–263

Kraft P (2011) Population stratification bias more widespread than previously thought. Epidemiology 22:408–409

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ (2007) Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63:111–119

Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7:385–394

Lake SL, Laird NM (2004) Tests of gene-environment interaction for case-parent triads with general environmental exposures. Ann Hum Genet 68:55–64

Lee WC, Chang CH (2006) Assessing effects of disease genes and gene-environment interactions: the case-spouse design and the counterfactual-control analysis. J Epidemiol Community Health 60:683–685

Lehr T, Yuan J, Zeumer D, Jayadev S, Ritchie MD (2011) Rule based classifier for the analysis of gene–gene and gene-environment interactions in genetic association studies. BioData Min 4:4

Lesch KP (2004) Gene-environment interaction and the genetics of depression. J Psychiatry Neurosci 29:174–184

Lettre G, Lange C, Hirschhorn JN (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol 31:358–362

Li D, Conti DV (2009) Detecting gene-environment interactions using a combined case-only and case-control approach. Am J Epidemiol 169:497–504

Lim S, Beyene J, Greenwood CM (2005) Continuous covariates in genetic association studies of case-parent triads: gene and gene-environment interaction effects, population stratification, and power analysis. Stat Appl Genet Mol Biol 4:Article20

Lindstrom S, Yen YC, Spiegelman D, Kraft P (2009) The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. Hum Hered 68:171–181

Little R, Rubin D (1987) Statistical analysis with missing data. Wiley, New York

Liu X, Fallin MD, Kao WH (2004) Genetic dissection methods: designs used for tests of gene-environment interaction. Curr Opin Genet Dev 14:241–245

Lo CY, Hsieh PH, Chen HF, Su HM (2009) A maternal high-fat diet during pregnancy in rats results in a greater risk of carcinogen-induced mammary tumors in the female offspring than exposure to a high-fat diet in postnatal life. Int J Cancer 125:767–773

Lobach I, Mallick B, Carroll RJ (2011) Semiparametric Bayesian analysis of gene-environment interactions with error in measurement of environmental covariates and missing genetic data. Stat Interface 4:305–316

Lou XY, Chen GB, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD (2008) A combinatorial approach to detecting gene–gene and gene-environment interactions in family studies. Am J Hum Genet 83:457–467

Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P (2004) Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 5:32

Macgregor S, Khan IA (2006) GAIA: an easy-to-use web-based application for interaction analysis of case-control data. BMC Med Genet 7:34

Maenner MJ, Denlinger LC, Langton A, Meyers KJ, Engelman CD, Skinner HG (2009) Detecting gene-by-smoking interactions in a genome-wide association study of early-onset coronary heart disease using random forests. BMC Proc 3(Suppl 7):S88

Mahachie John JM, Van Lishout F, Van Steen K (2011) Model-based multifactor dimensionality reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. Eur J Hum Genet 19:696–703

Maity A, Carroll RJ, Mammen E, Chatterjee N (2009) Testing in semiparametric models with interaction, with applications to gene-environment interactions. J R Stat Soc Series B Stat Methodol 71:75–96

Manning AK, LaValley M, Liu CT, Rice K, An P, Liu Y, Miljkovic I, Rasmussen-Torvik L, Harris TB, Province MA, Borecki IB, Florez JC, Meigs JB, Cupples LA, Dupuis J (2011) Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP × environment regression coefficients. Genet Epidemiol 35:11–18

Manolio TA, Collins FS (2007) Genes, environment, health, and disease: facing up to complexity. Hum Hered 63:63–66

Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417

McKinney BA, Crowe JE, Guo J, Tian D (2009) Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. PLoS Genet 5:e1000432

Meyer UA (2000) Pharmacogenetics and adverse drug reactions. Lancet 356:1667–1671

Mi X, Eskridge KM, George V, Wang D (2011) Structural equation modeling of gene-environment interactions in coronary heart disease. Ann Hum Genet 75:255–265

Moerkerke B, Vansteelandt S, Lange C (2010) A doubly robust test for gene-environment interaction in family-based studies of affected offspring. Biostatistics 11:213–225

Motsinger AA, Dudek SM, Hahn LW, Ritchie MD (2006) Comparison of neural network optimization approaches for studies of human genetics. Lect Notes Comput Sci 3907:103–114

Mukherjee B, Chatterjee N (2008) Exploiting gene-environment independence for analysis of case-control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64:685–694

Mukherjee B, Zhang L, Ghosh M, Sinha S (2007) Semiparametric Bayesian analysis of case-control data under conditional gene-environment independence. Biometrics 63:834–844

Mukherjee B, Ahn J, Gruber SB, Ghosh M, Chatterjee N (2010) Case-control studies of gene-environment interaction: Bayesian design and analysis. Biometrics 66:934–948

Mukherjee B, Ahn J, Gruber SB, Chatterjee N (2011) Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. Am J Epidemiol 175:177–190

Murcray CE, Lewinger JP, Gauderman WJ (2009) Gene-environment interaction in genome-wide association studies. Am J Epidemiol 169:219–226

Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ (2011) Sample size requirements to detect gene-environment interactions in genome-wide association studies. Genet Epidemiol 35:201–210

Ober C, Vercelli D (2011) Gene-environment interactions in human disease: nuisance or opportunity? Trends Genet 27:107–115

Paré G, Cook NR, Ridker PM, Chasman DI (2010) On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS Genet 6:e1000981

Park MY, Hastie T (2008) Penalized logistic regression for detecting gene interactions. Biostatistics 9:30–50

Pattin KA, White BC, Barney N, Gui J, Nelson HH, Kelsey KT, Andrew AS, Karagas MR, Moore JH (2009) A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. Genet Epidemiol 33:87–94

Pearce N (2011) Epidemiology in a changing world: variation, causation and ubiquitous risk factors. Int J Epidemiol 40:503–512

Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JP (2009) Discovery properties of genome-wide association signals from cumulatively combined data sets. Am J Epidemiol 170:1197–1206

Pereira TV, Patsopoulos NA, Pereira AC, Krieger JE (2011) Strategies for genetic model specification in the screening of genome-wide meta-analysis signals for further replication. Int J Epidemiol 40:457–469

Phillips PC (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9:855–867

Piegorsch WW, Weinberg CR, Taylor JA (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13:153–162

Prentice RL (2011) Empirical evaluation of gene and environment interactions: methods and potential. J Natl Cancer Inst 103:1209–1210

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463

Rappaport SM, Smith MT (2010) Environment and disease risks. Science 330:460–461

Ripley B (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge

Risch N, Herrell R, Lehner T, Liang KY, Eaves L, Hoh J, Griem A, Kovacs M, Ott J, Merikangas KR (2009) Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. JAMA 301:2462–2471

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147

Ritchie MD, Motsinger AA, Bush WS, Coffey CS, Moore JH (2007) Genetic programming neural networks: a powerful bioinformatics tool for human genetics. Appl Soft Comput 7:471–479

Rothman K, Greenland S (1998) Modern epidemiology. Lippencott-Raven, Philadelphia

Rothman KJ, Greenland S, Walker AM (1980) Concepts of interaction. Am J Epidemiol 112:467–470

Rothman K, Greenland S, Lash T (2008) Modern epidemiology, 3rd edn. Lippincott Williams & Wilkins, Philadephia

Schaid DJ (1999) Case-parents design for gene-environment interaction. Genet Epidemiol 16:261–273

Schwarz DF, Konig IR, Ziegler A (2010) On safari to random jungle: a fast implementation of random forests for high-dimensional data. Bioinformatics 26:1752–1758

Schwender H, Ruczinski I (2010) Logic regression and its extensions. Adv Genet 72:25–45

Shi M, Umbach DM, Weinberg CR (2011) Family-based gene-by-environment interaction studies revelations and remedies. Epidemiology 22:400–407

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

Smith GD, Timpson N, Ebrahim S (2008) Strengthening causal inference in cardiovascular epidemiology through Mendelian randomization. Ann Med 40:524–541

Song YS, Wang F, Slatkin M (2010) General epistatic models of the risk of complex diseases. Genetics 186:1467–1473

Stern MC, Johnson LR, Bell DA, Taylor JA (2002) XPD codon 751 polymorphism, metabolism genes, smoking, and bladder cancer risk. Cancer Epidemiol Biomark Prev 11:1004–1011

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8:25

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinformatics 9:307

Struchalin MV, Dehghan A, Witteman JC, van Duijn C, Aulchenko YS (2010) Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. BMC Genet 11:92

Takeuchi F, Kobayashi S, Ogihara T, Fujioka A, Kato N (2011) Detection of common single nucleotide polymorphisms synthesizing quantitative trait association of rarer causal variants. Genome Res 21:1122–1130

Tan P, Steinbach M, Kumar V (2006) Introduction to Data Mining. Addison-Wesley, Reading

Tan Y-D, Fornage M, George V. Xu H (2007) Parent–child pair design for detecting gene–environment interactions in complex diseases. Hum Genet 121:745–757

Tanck MW, Jukema JW, Zwinderman AH (2006) Simultaneous estimation of gene–gene and gene-environment interactions for numerous loci using double penalized log-likelihood. Genet Epidemiol 30:645–651

Tchetgen Tchetgen EJ, Robins J (2010) The semiparametric case-only estimator. Biometrics 66:1138–1144

Tchetgen Tchetgen EJ, Kraft P (2011) On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. Epidemiology 22:257–261

Tchetgen Tchetgen EJ, VanderWeele TJ (2012) Robustness of measures of interaction to unmeasured confounding. Harvard University Biostatistics Working Paper Series Working Paper 89

Thomas DC (2000) Case-parents design for gene-environment interaction by Schaid. Genet Epidemiol 19:461–463

Thomas D (2010a) Gene–environment-wide association studies: emerging approaches. Nat Rev Genet 11:259–272

Thomas D (2010b) Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Public Health 31:21–36

Thompson WD (1991) Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 44:221–232

Tryon R (1939) Cluster analysis. McGraw-Hill, New-York

Tung L, Gordon D, Finch SJ (2007) The impact of genotype misclassification errors on the power to detect a gene-environment interaction using cox proportional hazards modeling. Hum Hered 63:101–110

Tweel I, Schipper M (2004) Sequential tests for gene-environment interactions in matched case-control studies. Stat Med 23:3755–3771

Tzeng JY, Zhang DW, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF (2011) Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. Am J Hum Genet 89:277–288

Uher R (2008) Gene-environment interaction: overcoming methodological challenges. Novartis Found Symp 293:13–26 (discussion 26–30, 68–70)

Umbach DM, Weinberg CR (1997) Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 16:1731–1743

Umbach DM, Weinberg CR (2000) The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 66:251–261

van der Sluis S, Dolan CV, Neale MC, Posthuma D (2008) A general test for gene-environment interaction in sib pair-based association analysis of quantitative traits. Behav Genet 38:372–389

Van Lishout F, Cattaert T, Mahachie John M, Gusareva E, Urrea V, Cleynen I, Théatre E, Charloteaux B, Calle M, Wehenkel L, Van Steen K (2011) An efficient algorithm to perform multiple testing in epistasis screening

Van Steen K (2012) Travelling the world of gene–gene interactions. Brief Bioinform 13:1–19

Vansteelandt S, Demeo DL, Lasky-Su J, Smoller JW, Murphy AJ, McQueen M, Schneiter K, Celedon JC, Weiss ST, Silverman EK, Lange C (2008) Testing and estimating gene-environment interactions in family-based association studies. Biometrics 64:458–467

Vercelli D (2010) Gene-environment interactions in asthma and allergy: the end of the beginning? Curr Opin Allergy Clin Immunol 10:145–148

Visscher PM, Yang J, Goddard ME (2010) A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). Twin Res Hum Genet 13:517–524

Wakefield J, De Vocht F, Hung RJ (2010) Bayesian mixture modeling of gene-environment and gene–gene interactions. Genet Epidemiol 34:16–25

Wang LY, Lee WC (2008) Population stratification bias in the case-only study for gene-environment interactions. Am J Epidemiol 168:197–201

Wang T, Ho G, Ye K, Strickler H, Elston RC (2009) A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. Genet Epidemiol 33(1):6–15

Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118

Weinberg CR, Umbach DM (2000) Choosing a retrospective design to assess joint genetic and environmental contributions to risk. Am J Epidemiol 152:197–203

Whittemore AS (2007) Assessing environmental modifiers of disease risk associated with rare mutations. Hum Hered 63:134–143

Wild CP (2005) Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomark Prev 14:1847–1850

Willis-Owen SA, Valdar W (2009) Deciphering gene-environment interactions through mouse models of allergic asthma. J Allergy Clin Immunol 123:14–23 (quiz 24–25)

Witte JS, Gauderman WJ, Thomas DC (1999) Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol 149:693–705

Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ (2003) The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? Int J Epidemiol 32:51–57

Wong MY, Day NE, Luan JA, Wareham NJ (2004) Estimation of magnitude in gene-environment interactions in the presence of measurement error. Stat Med 23:987–998

Wray NR, Purcell SM, Visscher PM (2011) Synthetic associations created by rare variants do not explain most GWAS results. Plos Biol 9:e1000579

Wright AF, Carothers AD, Campbell H (2002) Gene-environment interactions–the BioBank UK study. Pharmacogenomics J 2:75–82

Wu X, Jin L, Xiong M (2009) Mutual information for testing gene-environment interaction. PLoS One 4:e4578

Wu C, Hu Z, He Z, Jia W, Wang F, Zhou Y, Liu Z, Zhan Q, Liu Y, Yu D, Zhai K, Chang J, Qiao Y, Jin G, Liu Z, Shen Y, Guo C, Fu J, Miao X, Tan W, Shen H, Ke Y, Zeng Y, Wu T, Lin D (2011) Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. Nat Genet 43:679–684

Wyszynski DF, Diehl SR (2001) The mother-only method (MOM) to detect maternal gene–environment interactions. Paediatr Perinat Epidemiol 15:317–318

Yang Q, Khoury MJ (1997) Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. Epidemiol Rev 19:33–43

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Yoshida M, Koike A (2011) SNPInterForest: a new method for detecting epistatic interactions. BMC Bioinformatics 12:469

Yu K, Wacholder S, Wheeler W, Wang Z, Caporaso N, Landi MT, Liang F (2012) A flexible bayesian model for studying gene-environment interaction. PLoS Genet 8:e1002482

Zhai R, Zhao Y, Liu G, Ter-Minassian M, Wu IC, Wang Z, Su L, Asomaning K, Chen F, Kulke MH, Lin X, Heist RS, Wain JC, Christiani DC (2011) Interactions between environmental factors and polymorphisms in angiogenesis pathway genes in esophageal adenocarcinoma risk: a case-only study. Cancer 118:804–811

Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39:1167–1173

Zhang L, Mukherjee B, Ghosh M, Gruber S, Moreno V (2008) Accounting for error due to misclassification of exposures in case-control studies of gene-environment interaction. Stat Med 27:2756–2783

Zhang Y, Jiang B, Zhu J, Liu JS (2011) Bayesian models for detecting epistatic interactions from genetic data. Ann Hum Genet 75:183–193

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci USA 109:1193–1198