Data-driven integration of epidemiological and toxicological data to select candidate interacting genes and environmental factors in association with disease

Chirag J. Patel[1], Rong Chen[1,2] and Atul J. Butte[1,2],*

[1]Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305, USA and [2]Lucile Packard Children's Hospital, Palo Alto, CA, 94304, USA

ABSTRACT

Motivation: Complex diseases, such as Type 2 Diabetes Mellitus (T2D), result from the interplay of both environmental and genetic factors. However, most studies investigate either the genetics or the environment and there are a few that study their possible interaction in context of disease. One key challenge in documenting interactions between genes and environment includes choosing which of each to test jointly. Here, we attempt to address this challenge through a data-driven integration of epidemiological and toxicological studies. Specifically, we derive lists of candidate interacting genetic and environmental factors by integrating findings from genome-wide and environment-wide association studies. Next, we search for evidence of toxicological relationships between these genetic and environmental factors that may have an etiological role in the disease. We illustrate our method by selecting candidate interacting factors for T2D.

Contact: abutte@stanford.edu

1 INTRODUCTION

Complex diseases and phenotypes, such as Type 2 Diabetes (T2D), have multifactorial etiology in which both genetic and environmental factors play a role (Schwartz and Collins, 2007). Through multiple genetic association studies, such as the genome-wide association study (GWAS) we have discovered many common variants associated with complex disease; however, these variants confer very little disease risk and cumulatively explain a limited portion of heritability (Manolio *et al.*, 2009). It is hypothesized that perhaps some of the unexplained risk may be due to 'gene-environment' interactions, or joint effect of a genetic and environmental factor are different than that of each of the factors alone (Hunter, 2005).

We recently put forth an analogous framework to investigate possible environmental influences on disease, called an 'environment-wide association study' (EWAS), where we connected over 250 environmental factors to T2D and serum cholesterol levels (Patel *et al.*, 2010; Patel *et al.*, 2012). Humans are not exposed to a few environmental factors at a time, but many types of infectious agents, pollutants, nutrients and vitamins at once, all of which may play a role in disease. EWAS evaluates multiple environmental factors to capture the complex nature of environmental exposure in relation to disease.

Nevertheless, GWAS and EWAS only examine genetic or environmental factors, respectively, without consideration of the other. Examining interactions is complex, power-intensive and exacts a multiple comparison burden (Thomas, 2010). For example, a list of *G* genetic variants and *E* environmental exposures would result in *G* or *E* hypothesis tests in GWAS and EWAS, respectively; however, to screen the possible space of interactions would many more tests, equal to *G* times *E*. To accommodate these challenges, statistical methods have been developed to larger space of interactions (Thomas, 2010; Thomas *et al.*, 2011). Another option includes adapting methods used to detect gene–gene interactions, or epistasis (Cordell, 2009).

Yet another way includes paring down the list of *G* and *E* factors to test. In many variant by environment investigations, factors are selected by convenience, without sufficient documentation of the strength of their marginal associations. There is a need for methods to select common variants and exposures resulting from comprehensive studies, such as GWAS and EWAS (Grarup and Andersen, 2007; Khoury and Wacholder, 2009; Omenn, 2010). Further still, GWAS and EWAS operate on the population scale, and there is need to integrate molecular-scale toxicological evidence—such as how an environmental factor might modulate a biological process—between exposures and genes. As yet, there are few ways to integrate environmental findings from epidemiological and toxicological studies.

Here, we propose integrating findings from genetic and environmental association studies with knowledge regarding molecular- and toxicological-scale interactions. By integrating these data, we aim to assemble a list of genetic and environmental factors to later test jointly against a phenotype or disease of interest.

Specifically, we attained a comprehensive list of genetic variants associated with disease from a curated database, VARiants Informing MEDicine (VARIMED). VARIMED contains results from over 6500 peer-reviewed genetic epidemiology studies (e.g. GWAS), covering over 110 000 variants in 9700 genes associated with over 2000 phenotypes (Chen *et al.*, 2010). To prioritize environmental factors, we used an EWAS approach with the Health and Nutrition Examination Survey (NHANES), a biannual health survey whose participants are representative of the US population [Centers for Disease Control and Prevention (CDC), 2009].

Last, we connected these findings from VARIMED and EWAS with the Comparative Toxicogenomics Database (CTD), a database of curated findings in toxicology. Toxicology is the study of the biological response to chemicals, such as how a chemical exposure might induce a gene expression change. As an example, one such toxicological interaction is derived from an investigation studying gestational exposure of the plasticizing agent Bisphenol A on adipogenesis in a *Rattus norvegicus* model system (Somm *et al.*, 2009), in which mRNA levels of the gene *LPL* increased

---

*To whom correspondence should be addressed.

after chemical exposure. In the CTD, many such chemical–gene relationships are curated from biomedical journals. The CTD has documented over 200 000 of these gene and environment relationships spanning over 26 000 publications. We show how one can use genetic and environmental epidemiological results with the CTD to create lists of candidate interacting genes, variants and environmental factors that may be indicative of disease etiology.

2   METHODS

We propose integrating genetic and environmental epidemiological findings with molecular toxicological data to create a list of interacting candidate factors. A schematic of our methodology is depicted in Figure 1. First, we describe the data sources we integrated.

2.1   VARIMED: a comprehensive disease-SNP association database

VARIMED is a database of human variants associated with disease and disease-related phenotypes curated from the biomedical literature (Fig. 1A; Chen *et al.*, 2010). As described, we downloaded all abstracts from MEDLINE and identified published manuscripts that have reported on findings related to human disease and genetic association through their medical subject headings (MEsH; http://www.ncbi.nlm.nih.gov/mesh). Specifically, we searched for terms such as 'Genetic Variation', 'Genetic predisposition to disease', among other terms. We also downloaded all papers from publicly-available disease-variant databases, such as the National Human Genome Research Institute's GWAS catalog (Hindorff *et al.*, 2009). Features captured from these papers including the disease or phenotype under investigation (e.g. T2D or cholesterol levels), the race of the population (e.g. European and African), study type (e.g. case-control or cohort), and most importantly, variant identifier (dbSNP), *P*-value of association and effect size (e.g. odds ratio). In its current form, this database covers 6890 publications, 111 977 variants mapped to 9752 genes and 2053 phenotypes. Further, each variant was mapped to a gene (e.g. intron, exon or promoter region), utilizing the National Centers of Biotechnology Information's dbSNP resource (human build 131; National Center for Biotechnology Information, 2012).

2.2   Comparative toxicogenomics database

The CTD includes manually-curated, cross-species relations between chemicals and genes, proteins and mRNA transcripts (Fig. 1D; Mattingly *et al.*, 2006). We downloaded the database (January 2012), spanning 219 618 direct relationships between 6079 unique environmental chemical factors and 21 267 genes and their products in 336 organisms, including humans. An example of a relationship in the CTD is 'Chemical γ-tocopherol results in increased expression of *PPARG* mRNA as cited by Campbell *et al.* in *Homo sapiens* cell lines'.

2.3   National health and nutrition examination survey

The NHANES is a now biannual cross-sectional health survey representative of the US population in progress since 1999 [Fig. 1B; Centers for Disease Control and Prevention (CDC), 2009]. Each biannual survey is an independent, non-overlapping sample of 10 000 participants. Roughly a third of these participants have their serum and urine analyzed for environmental factors. In total, there are 266 unique measures of environmental factors, including infectious agents (e.g. Hepatitis A and HIV), chemical pollutants (e.g. pesticides and hydrocarbons) and nutrients (e.g. vitamins A, B, C, D, E and carotenoids). Most environmental factor measures were continuous; however, some factors were binary, such as presence/absence assays for some infectious agents. Further, these participants have clinical measures, such as serum cholesterol and fasting blood glucose (FBG), markers for chronic disease. Here, we downloaded all available NHANES

laboratory data for 1999–2000, 2001–2002, 2003–2004 and 2005–2006 surveys. To facilitate later integration with the CTD, we mapped each environmental factor with a MeSH identifier [e.g. Bisphenol A is C006780 (http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&term=C006780)].

2.4   The space of potential molecular interactions between phenotype-associated genes and environmental factors

Before integrating EWAS findings with VARIMED and toxicological data to build a list of potential interactions related to the etiology of disease, we sought to observe how many documented molecular relationships exist between phenotype- or disease-associated genes and environmental factors in general. If many of such relationships exist, we concluded, such a database would be adequate for the next step (Fig. 1C–E). Specifically, we extracted all associations in which a variant was associated to any phenotype (e.g. HDL Cholesterol, T2D and Hypertension) with a *P*-value $\leq 1 \times 10^{-6}$ and mapped these variants to genes. We observed how many of these genes (or their products) have a documented molecular interaction with a factor in the CTD.

2.5   Integration of epidemiological findings with toxicological data to prioritize factors

We propose combining epidemiological results from studies such as GWAS and as EWAS with toxicological molecular data to attain a candidate list of interacting genetic variants and environmental factors for a particular disease (Fig. 1). We illustrate this process using T2D on populations of European ancestry.

In the first step, we attain a comprehensive list of genetic variants associated with T2D from VARIMED in populations of European ancestry (Fig. 1B, left panel). We choose a set of *G* variants based on their strength of association, or *P*-value set at some threshold *p*.

In the second step, we attain a similar list of environmental factors associated with T2D with the EWAS approach in an analogous population in NHANES (Fig. 1B, right panel; Patel *et al.*, 2010). We briefly describe this approach. First, we classified diabetics and controls as individuals who had FBG $\geq$126 mg/dL, the recommended threshold for diagnosis of T2D from the American Diabetes Association. Second, we filtered these participants by self-report race, only considering individuals in the Non-Hispanic White group. Third, we filtered out 7 environmental exposure variables that varied little and were thus uninformative, leaving 244 variables for analysis. Most continuous environmental exposure variables had a long-tailed distribution and we log-transformed and applied a *z*-transformation to them (mean centered and divided by the SD). Due to elimination of all but Non-Hispanic White participants from the sample, we combined individuals from all surveys to maximize power for detection.

We conducted our EWAS by associating each of the 244 factors to T2D status using multivariate survey logistic regression, adjusting for sex, age, socioeconomic level (SES) and body mass index (BMI) iteratively, resulting in 244 separate tests of association, indexed by *i*:

$$\text{Logit}(\text{FBG} \geq 126\text{mg/dL}) = \alpha + \beta_{\text{exp}}\text{exposure}_i + \text{other adjustments}$$

We collected the effect size of each *i*-th exposure (β_{exp}) and its *P*-value. Each test had a sample size of 315–3748 (comprising of 8–10% diabetics). We chose to adjust for age, sex, socioeconomic level and BMI as they are possible confounders and can be correlated with both T2D status and environmental exposure (Patel *et al.*, 2010). We accounted for multiple comparisons and assessed strength of association by estimating the false discovery rate (FDR), the estimated proportion of false discoveries made versus the number of discoveries made at a given significance level (Benjamini and Hochberg, 1995) and as we described previously (Patel *et al.*, 2010). To estimate the FDR, we first permuted the dependent variable, T2D status, 1000 times and re-computed regressions to attain a distribution of 'null' *P*-values by which we could estimate the average number of false
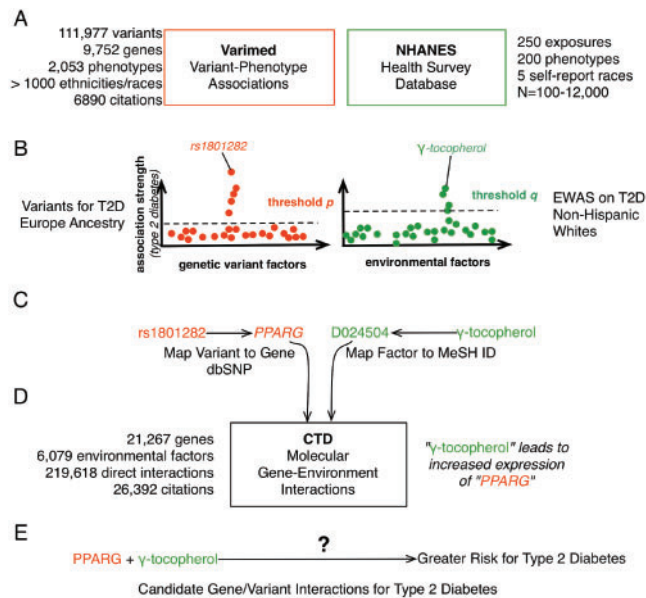
Fig. 2. Proportion of VARIMED genes covered in CTD versus number of total genes associated for a particular phenotype. We associated a gene with a phenotype bases on presence of a variant in an intron, exon or promoter region for that gene. Variants are associated with a phenotype with a P-value $<1 \times 10^{-6}$. Phenotypes with >20 genes, such as 'Metabolic Trait', 'Height' and 'Rheumatoid arthritis' are labeled with numbers (key on right panel). T2D is depicted as 17, with 100% of 43 genes having some documented molecular interaction in the CTD. Values on the x-axis jittered to show density

Fig. 1. Method of integrating epidemiological and toxicological data to create variant by environment interaction candidates. (**A**) Summary of epidemiological databases, VARIMED (source of genetic associations, red) and NHANES (source of environmental data, green). (**B**) Factors are chosen by their strength of association to T2D depicted by Manhattan plots from VARIMED or EWAS process with NHANES. A threshold p and q are chosen to choose candidate factors (examples in italics). (**C**) Variants are mapped to gene promoter, exon or intron; environmental factors are mapped to MeSH ID. (**D**) Factors are queried for curated evidence regarding molecular-level interaction in CTD. Knowledge is coded in CTD in an example on right. (**E**) Presence of interacting pair in CTD is a candidate for further study in context of T2D

positives for a given threshold. Then, we estimated the FDR as the rate of false positives divided by discoveries at a given P-value. Finally, we choose a set of E exposures based on their strength of association at some FDR $\leq q$.

Given a set of E exposures selected at some threshold q and G genes selected with a threshold p, we search for any pair from E and G ($E \times G$ total pairs possible) in the CTD (Fig. 1D, E). Finally, the disease-associated variants that map to genes may be utilized to as candidate variants for a variant by environment study.

We implemented a permutation method to approximate the probability of attaining such a result due to random chance. Specifically, we permuted the dependent variable of EWAS analysis (T2D status) as described in the FDR estimation procedure above and recomputed the association between exposure and T2D. We took the top E′ (equal to the size of E, |E|) exposures and searched for any pair E′ and G in the CTD. We then assessed whether any or all pairs from E and G are in randomization-derived set, constructed with E′ and G. We repeated this process 1000 times and computed the frequency that any or all pairs from E and G appear in E′ and G. This frequency was our estimate of the empirical P-value.

3 RESULTS

3.1 Baseline space of molecular interactions

We first sought to observe if the CTD contained toxicological relationships between environmental chemicals and genes with variants identified in association studies (Fig. 2). We found that most genes associated with disease have some documented molecular
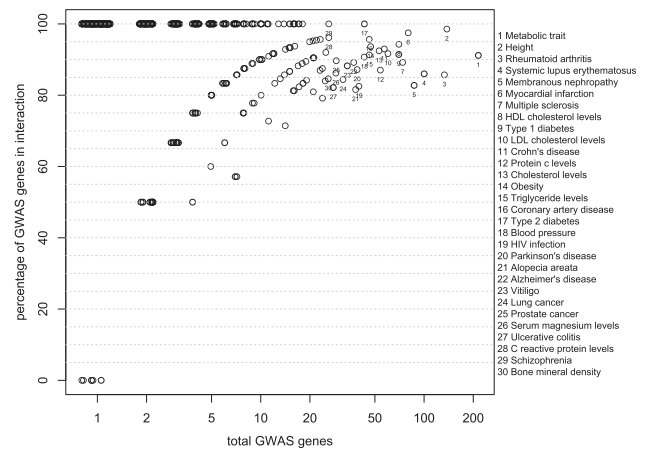
interactions with chemical environmental factors. First, in total, we tabulated 2236 unique genes associated that mapped to a variant with P-value $\leq 1 \times 10^{-6}$ and all of these 2236 genes had some documented relationship in the CTD. Second, we found that many of the genes associated with specific phenotypes had adequate coverage as well. For example, there were 113 variants associated with T2D (P-value $\leq 1 \times 10^{-6}$) in 43 genes among 11 different populations.

Of these 43 genes, all 43 had at least one documented molecular interaction with an environmental factor as curated in the CTD (Fig. 2, number point labeled 17). In fact, for most phenotypes, >40% of genes with genome-wide significant variants had a documented interaction in the CTD. Of the 508 phenotypes with significantly associated genes, 392 had at least 90% of their genes covered in CTD (Fig. 2). However, 182 phenotypes had but only 1 significantly associated gene in VARIMED and 176 of these genes had coverage in the CTD. We concluded that the CTD had adequate coverage for many common, complex disorders.

3.2 Integration of epidemiological findings with toxicological data to prioritize candidate factors for T2D

In our EWAS process, we found 4, 9, 28 and 68% unique environmental factors associated with T2D at FDR, or q, of 5, 10, 20 and 30%, respectively, in the Non-Hispanic White population. For example, at q <10%, we found factors related to T2D such as γ-tocopherol, a type of vitamin E (q = 1%), Heptachlor Epoxide, a pesticide (q = 6%) with effect sizes of 1.4 and 1.8, respectively, for a change in 1 SD in logged units of exposure level. We found other factors protective for T2D, such *cis-β*-carotene (q = 1%), *trans-β*-carotene (q = 1%) and vitamin D (q = 1%) with odds ratios of 0.66, 0.70 and, respectively, for a change in 1 SD in logged units of exposure. As we expected, these findings were in line our previously published findings where we analyzed all races available
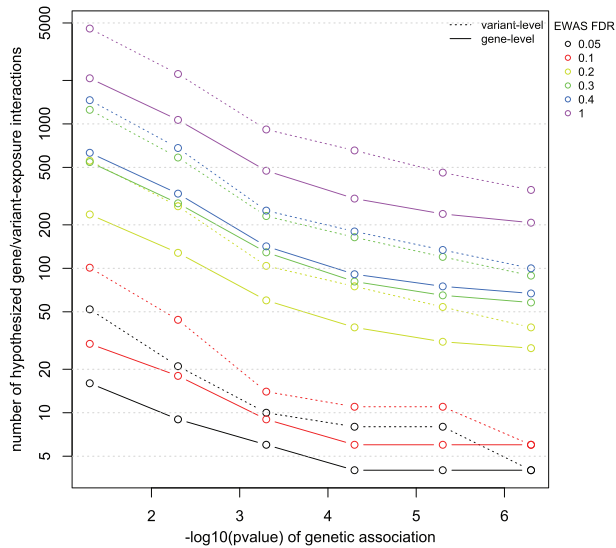
Fig. 3. Number of candidate genetic factor by environmental factor interactions derived from varying significance thresholds from genetic association (p, x-axis) and EWAS FDR (or q, in different colors). Solid lines depict total gene-by-environment interactions. Dotted lines depict number of variant-by-environment interactions. For example, there are six interacting pairs in the CTD for genes and exposures found at a $p = 5 \times 10^{-7}$ and $q = 10\%$, respectively

(Patel *et al.*, 2010). In that study, we were able to validate in more than one survey γ-tocopherol, Heptachlor Epoxide, β-carotenes and a related PCB isomer, with a discovery FDR of 10% (Patel *et al.*, 2010).

We queried VARIMED for variants associated with T2D in populations of European ancestry and mapped these variants to introns, exons and promoters of genes. For example, we curated 690, 170 and 75 variants associated with T2D with P-values 0.05, 5×10^{-4} and 5×10^{-6}, respectively. These variants mapped to 359, 85 and 35 unique genes, respectively.

Next, we looked for evidence of molecular interaction in CTD between factors found in EWAS and VARIMED for varying significance thresholds p and q (Fig. 3). As one would expect, for relaxed p and q, one would have molecular evidence in the CTD for over 2070 pairs of genes and environmental factors. This number is more than doubled when considering all disease variants within

those genes (4574). When prioritizing factors to p and q at levels close to EWAS and GWAS level of significance (5×10^{-7} and 10% FDR, respectively), one is left with only six candidate interacting pairs.

As an example, we show 6 candidate interacting factors produced from applying significance thresholds corresponding to p threshold of 5×10^{-6} (mapping to 35 genes) and q of 10% (9 environmental factors from EWAS; Table 1). Of these possibilities, there is evidence in the CTD that the favorable environmental factors, β-carotene and vitamin D may decrease gene expression (as referenced in Callens *et al.*, 2010; Kawada *et al.*, 2000), whereas the γ-tocopherol—associated with T2D risk—may increase *PPARG* gene expression (Campbell *et al.*, 2003).

The set of six interacting candidates did not appear in any randomization trials (empirical P-value <0.001); however, having at least one interaction had a P-value of 0.15.

Scripts and data for our procedure may be found at the following URL: http://www.stanford.edu/~cjp/t2d_varimed_ewas_ctd.

DISCUSSION

We have described a method to combine epidemiological association findings with results from toxicological experiments to create a tractable list candidate of genes, variants and environmental factors that might interplay in the context of disease. Additionally, the method is a data-driven integration of three disparate datasets that span the population to toxicological scale.

Of interest, this method has drawn attention to nutrient factors that may interact with a gene known to be functional and have a therapeutic role in T2D, *PPARG* (Murphy and Holder, 2000; Spiegelman, 1998). γ-Tocopherol, associated with T2D risk in EWAS, has been observed to increase expression of *PPARG* mRNA. Further, forms of carotene protective for T2D have an opposite role, reducing *PPARG* expression. Of note, these candidate interactions are derived from two very different toxicological investigations, one studying the influence of γ-tocopherol on *PPARG* expression in human cell lines (Campbell *et al.*, 2003) and the other observing the effects of carotenoids on *PPARG* in adipocyte differentiation in *Mus musculus* (Kawada *et al.*, 2000).

A next possible step includes testing these candidate interactions in the context of disease. Specifically, we propose an epidemiological study in which the interaction between functional variants in *PPARG*, such as rs1801282, be examined along with

Table 1. Predicted variant by environment interactions for factors prioritized by strength of genetic or environmental association to T2D

E-factor	*E*-factor OR (95% CI)	Gene (snp) (reference)	Variant OR (95% CI)	CTD-curated molecular interaction (reference)
γ-Tocopherol	1.38 (1.20–1.60)	PPARG [rs1801282] (Zeggini *et al.*, 2007)	1.14 (1.08–1.2)	Increases gene expression (Campbell *et al.*, 2003)
Cis-β-carotene	0.66 (0.55–0.79)	PPARG [rs1801282] (Zeggini *et al.*, 2007)	1.14 (1.08–1.2)	Decreases gene expression (Kawada *et al.*, 2000)
Vitamin D	0.68 (0.57–0.81)	PPARG [rs1801282] (Zeggini *et al.*, 2007)	1.14 (1.08–1.2)	Affects gene expression (Callens *et al.*, 2010)
Trans-β-carotene	0.70 (0.58–0.82)	PPARG [rs1801282] (Zeggini *et al.*, 2007)	1.07 (1.05–1.09)	Decreases gene expression (Kawada *et al.*, 2000)
PCB187	1.89 (1.3–2.8)	PRC1 [rs8042680] (Voight *et al.*, 2010)	1.07 (1.05–1.09)	Increases gene expression (Buterin *et al.*, 2006)
Total β-carotene	0.71 (0.60–0.85)	PPARG [rs1801282] (Zeggini *et al.*, 2007)	1.14 (1.08–1.2)	Decreases gene expression (Kawada *et al.*, 2000)

Each row depicts a candidate interaction pair, its marginal effect size (Odds ratio) and type, and citation of molecular interaction (right column). For example, in the first row, γ-tocopherol has a documented molecular interaction with *PPARG* ('increasing gene expression'). Odds ratios for environmental factor are for 1 SD change in logged exposure variable. Variant odds ratios are for an additive genetic model. Factors chosen based on strength of marginal association ($p \le 5 \times 10^{-6}$ and $q \le 10\%$).

these nutrients. Specifically, we shall test whether different levels of these nutrients modulate genetic risk for T2D.

Variant or gene-by-environment interactions may occur among factors that fail to reach a suitable level of statistical significance association studies (p and q). Our method will not prioritize these factors. However, selecting which of the other many variants to test is challenging. It is well known that testing for interactions is power-intensive (Hunter, 2005) and testing a large number or all of them imposes an even greater power and multiplicity burden (Thomas, 2010). For environmental factors, the choice of which ones to test for interaction is even more vexing. In contrast to common genetic variants, there is yet no high-throughput measurement platform that captures all the environmental factors [the 'exposome' (Rappaport and Smith, 2010)] and lack of measuring capacity limits data availability. Measurement error and cost—unlike the genome—can be substantial for many environmental exposures (Ioannidis *et al.*, 2009; Vineis, 2004). Further, it is postulated that if a study is underpowered to detect main effects then one may be underpowered to detect the interaction effect (Hunter, 2005).

Furthermore, candidate interaction pairs are also influenced by the significance thresholds p and q. We chose thresholds that were close to genome-wide levels of significance and previously documented EWAS FDR threshold (Patel *et al.*, 2010). We show how the number of candidates varies as a function of p and q (Fig. 3), and the investigator may use this relationship to pick an optimal number of candidates to pursue for further testing based on their resources. While we attempt to compute the probability of a subset of such findings using randomization, we acknowledge that other ways gauging the expectation of such relationships such as should be developed.

EWAS has both advantages and disadvantages (Fallin and Kao, 2011; Porta *et al.*, 2012). By allowing for a transparent and data-driven search for environmental factors associated with disease, it bypasses the problem of selectively testing and reporting a few associations at a time, which has been debated as a source of biased results in environmental epidemiological studies (Blair *et al.*, 2009; Boffetta *et al.*, 2008; Ioannidis, 2005; Ioannidis *et al.*, 2009). Like GWAS, EWAS provides an opportunity to find novel factors and prioritize them for future study. EWAS attempts to capture the complex and multifactorial relationship between environmental exposure and disease. However, associations found by EWAS may be biased. First, confounding bias, where the exposure of interest is correlated with a yet another exposure and the disease remains a major issue. For example, it is well known that SES is correlated with to both disease risk and exposure to environmental factors. Thus, the association of an exposure to a disease may not be causal, but related to the confounding variable such as SES. Yet another factor is 'reverse causality' (Ioannidis *et al.*, 2009), whereby the exposure comes as a result of the disease. One example of this includes healthy eating behavior after disease diagnosis, leading to associations between nutrient factors and the disease. While we have attempted to minimize these biases through linear adjustement and sensitivity analyses as described previously (Patel *et al.*, 2010; Patel *et al.*, 2012), we cannot conclude that these factors are indeed causal.

On the other hand, presence of biological interactions between a genetic and environmental factor may be indicative of a causal relationship between the environmental factor and disease (Davey Smith, 2010). Here, we propose how to generate such interaction candidates. Validation of interaction of such candidates will strengthen the evidence for our EWAS-derived hypotheses.

The public availability of genetic association and representative environmental data may allow for quick adoption by investigators to create etiological interaction hypotheses using existing cohorts. However, there are some limitations to these data. First, the majority of GWAS have occurred among populations of Caucasian ancestry and there is need for GWAS in other ancestral groups. For environmental measures the NHANES is a unique resource, containing a large number of phenotypic and environmental information on a population representative of the USA. However, it is but one source, and we lack access to other populations. Further, the environmental measures selected by the Centers of Disease Control and Prevention reflect prevalent exposures, but they are only a subset of the 'exposome' (Wild, 2005). Last, the NHANES participants represent non-institutionalized individuals and less prevalent diseases with hypothesized environmental influences, such as T1D, are not characterized in these surveys. Other methods to select candidate exposures may be employed, such as mining the biomedical literature (Liu *et al.*, 2009; Schmidt-Kastner *et al.*, 2012).

As any other curated database, the information in the CTD may be biased toward genes or chemicals that are known to take part in some toxicological or biological response. For example, the most cited gene and gene product in the CTD includes the cytochrome P450 class of genes (e.g. *CYP1A1*), known to play a role in chemical metabolism. Similarly, 'endocrine disrupting' chemicals such as estrogens and dioxins are among most highly cited chemicals in the CTD.

Of course, one may manually peruse databases like CTD to come up with candidate lists of environmental factors that interact with genes of interest. However, by integrating EWAS, investigators can simultaneously identify prevalent environmental factors and quantify its main effect. Both exposure prevalence and degree of effect are critical factors for public health. In short, we have described a method to choose candidate interacting pairs that bridges population relevance with toxicological evidence.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B.*, **57**, 289–300.

Blair,A. *et al.* (2009) Epidemiology, public health, and the rhetoric of false positives. *Environ. Health Perspect.*, **117**, 1809–1813.

Boffetta,P. *et al.* (2008) False-positive results in cancer epidemiology: a plea for epistemological modesty. *J. Natl Cancer Inst.*, **100**, 988–995.

Buterin,T. *et al.* (2006) Convergent transcriptional profiles induced by endogenous estrogen and distinct xenoestrogens in breast cancer cells. *Carcinogenesis*, **27**, 1567–1578.

Callens,C. *et al.* (2010) Targeting iron homeostasis induces cellular differentiation and synergizes with differentiating agents in acute myeloid leukemia. *J. Exp. Med.*, **207**, 731–750.

Campbell,S.E. *et al.* (2003) Gamma (gamma) tocopherol upregulates peroxisome proliferator activated receptor (PPAR) gamma (gamma) expression in SW 480 human colon cancer cell lines. *BMC Cancer*, **3**, 25.

Centers for Disease Control and Prevention (CDC). (2009) *National Health and Nutrition Examination Survey*. CDC, Atlanta, GA.

Chen,R. *et al.* (2010) Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS ONE*, **5**, e13574.

Cordell,H.J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.

Davey Smith,G. (2010) Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. *Genes Nutr.*, **6**, 27–43.

Fallin,M.D. and Kao,W.H.L. (2011) Is 'X'-WAS the future for all of epidemiology? *Epidemiology*, **22**, 457–459.

Grarup,N. and Andersen,G. (2007) Gene-environment interactions in the pathogenesis of type 2 diabetes and metabolism. *Curr. Opin. Clin. Nutr. Metab. Care*, **10**, 420–426.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Hunter,D.J. (2005) Gene-environment interactions in human diseases. *Nat. Rev. Genet.*, **6**, 287–298.

Ioannidis,J.P.A. (2005) Why most published research findings are false. *PLoS Med.*, **2**, e124.

Ioannidis,J. *et al.* (2009) Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. *Sci. Transl. Med.*, **1**, 8.

Kawada,T. *et al.* (2000) Carotenoids and retinoids as suppressors on adipocyte differentiation via nuclear receptors. *Biofactors*, **13**, 103–109.

Khoury,M.J. and 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 234–225.

Liu,Y.I. *et al.* (2009) The 'etiome': identification and clustering of human disease etiological factors. *BMC Bioinform.*, **10** (Suppl. 2), S14.

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

Mattingly,C.J. *et al.* (2006) The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.*, **92**, 587–595.

Murphy,G.J. and Holder,J.C. (2000) PPAR-gamma agonists: therapeutic role in diabetes, inflammation and cancer. *Trends Pharmacol. Sci.*, **21**, 469–474.

National Center for Biotechnology Information. (2012) dbSNP. National Center for Biotechnology Information.

Omenn,G.S. (2010) Overview of the symposium on public health significance of genomics and eco-genetics. *Annu. Rev. Public Health*, **31**, 1–8.

Patel,C.J. *et al.* (2010) An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS ONE*, **5**, e10746.

Patel,C.J. *et al.* (2012) Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels. *Int. J. Epidemiol.* [Epub ahead of print, doi: 10.1093/ije/dys003].

Porta,M. *et al.* (2012) Commentary: a step towards more comprehensive analyses of life course effects of mixtures of environmental factors. *Int. J. Epidemiol.* [Epub ahead of print, doi: 10.1093/ije/dys014].

Rappaport,S.M. and Smith,M.T. (2010) Environment and disease risks. *Science*, **330**, 460–461.

Schmidt-Kastner,R. *et al.* (2012) An environmental analysis of genes associated with schizophrenia: hypoxia and vascular factors as interacting elements in the neurodevelopmental model. *Mol. Psychiatry*. [Epub ahead of print, doi: 10.1038/mp.2011.183].

Schwartz,D. and Collins,F. (2007) Medicine. Environmental biology and human disease. *Science*, **316**, 695–696.

Somm,E. *et al.* (2009) Perinatal exposure to bisphenol a alters early adipogenesis in the rat. *Environ. Health Perspect.*, **117**, 1549–1555.

Spiegelman,B.M. (1998) PPAR-gamma: adipogenic regulator and thiazolidinedione receptor. *Diabetes*, **47**, 507–514.

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

Thomas,D.C. *et al.* (2011) Invited commentary: GE-Whiz! ratcheting gene-environment studies up to the whole genome and the whole exposome. *Am. J. Epidemiol.* [Epub ahead of print, doi: 10.1093/aje/kwr365].

Vineis,P. (2004) A self-fulfilling prophecy: are we underestimating the role of the environment in gene-environment interaction research? *Int. J. Epidemiol.*, **33**, 945–946.

Wild,C.P. (2005) Complementing the genome with an 'exposome': the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol. Biomarkers Prev.*, **14**, 1847–1850.

Zeggini,E. *et al.* (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336–1341.