American Journal of Epidemiology
Vol. 186, No. 7
DOI: 10.1093/aje/kwx229
Advance Access publication:
July 15, 2017

Special Article

Incorporation of Biological Knowledge Into the Study of Gene-Environment Interactions

Marylyn D. Ritchie*, Joe R. Davis, Hugues Aschard, Alexis Battle, David Conti, Mengmeng Du, Eleazar Eskin, M. Daniele Fallin, Li Hsu, Peter Kraft, Jason H. Moore, Brandon L. Pierce, Stephanie A. Bien, Duncan C. Thomas, Peng Wei, and Stephen B. Montgomery*

* Correspondence to Dr. Stephen B. Montgomery, Departments of Genetics and Pathology, Stanford University School of Medicine, Stanford, CA 94305 (e-mail: smontgom@stanford.edu); or Dr. Marylyn D. Ritchie, Geisinger Health System, 205 Hood Center for Health Research, Center Street, Danville, PA 17821(e-mail: marylyn.ritchie@psu.edu).

Initially submitted November 7, 2016; accepted for publication April 10, 2017.

A growing knowledge base of genetic and environmental information has greatly enabled the study of disease risk factors. However, the computational complexity and statistical burden of testing all variants by all environments has required novel study designs and hypothesis-driven approaches. We discuss how incorporating biological knowledge from model organisms, functional genomics, and integrative approaches can empower the discovery of novel gene-environment interactions and discuss specific methodological considerations with each approach. We consider specific examples where the application of these approaches has uncovered effects of gene-environment interactions relevant to drug response and immunity, and we highlight how such improvements enable a greater understanding of the pathogenesis of disease and the realization of precision medicine.

data integration; functional genomics; gene-environment interaction; model organisms

Abbreviations: eQTL, expression quantitative trait locus; GWAS, genome-wide association study; G×E, gene-environment interaction; SNP, single nucleotide polymorphism.

In the quest for the discovery of genetic and environmental risk factors associated with common, complex disease risk, researchers have largely focused on either the genome or "the exposome," the multitude of environmental factors affecting the individual's health. To identify genetic risk factors, genetics researchers have used a variety of study design techniques, including family-based linkage studies, candidate-gene association studies, and more recently genome-wide association studies (GWASs). Through unbiased, genome-wide scans, GWASs have identified thousands of common genetic variants associated with risk for common diseases (1), some of which highlight new biological pathways. However, the effects of associated variants are typically small and account for only a small proportion of the estimated heritability (2, 3). To identify environmental risk factors, epidemiologists have applied diverse study designs ranging from observational studies of the environment—investigating factors outside of the control of the individual (air pollution, water contaminants, etc.)—to experimental studies of modifiable risks, such as nutritional interventions. Classical studies of single environmental factors, such as radon and smoking, have greatly advanced our understanding of health (4). However, in recent years, large-scale environmental screening projects have emerged to enable simultaneous study of multiple environmental factors, or the exposome. These large-scale investigations are subject to many of the challenges of "big data" research, including high correlation among study variables, multiple testing corrections, and missing data (5).

The risk of common diseases is often due to a complex interplay of the genome and the exposome; however, the degree to which genetic versus environmental factors alter risk varies greatly by disease. For example, in certain subtypes of lung cancer, carcinogens in tobacco smoke have such a strong influence on disease risk that the effects of genetic variation pale in comparison (6). Conversely, a familial subtype of breast cancer, resulting from mutations in *BRCA1*/*BRCA2* genes, is predominantly explained by the genetic component of risk (7). Given that individual variability in both genes and environment can influence disease risk, joint analysis of the genome and the exposome, as well as their potential interactions, will provide better insights into disease etiology. However, with tens of millions of variants and thousands of measured environmental variables, the challenge is identifying how to do so effectively.

Publicly funded research has provided a wealth of genetic and environmental data. The *All of Us* Research Program announced in 2015 aims to further support data-driven science by building a national research cohort of 1 million or more US participants with genetic and environmental data (https://www.nih.gov/research-training/allofus-research-program). As these types of resources grow, the challenges of big-data analysis are a significant consideration for genomic and exposome research. Analyzing data sets that include both types of data only exacerbate these issues further. GWASs have had remarkable success in combining analyses across diverse studies to improve power. However, for exposome studies, challenges remain in unifying measurements and defining analytical procedures. When such analyses can be achieved, several challenges emerge. These issues include computational complexity issues, a very large multiple-testing burden, big *p* small *n* (many more variables than individuals/samples), and often sparse data matrices (due to missing data). In order to deal with these issues, many approaches perform some form of data reduction or filtering prior to analysis to increase statistical power to detect interactions.

One strategy for data reduction is to use genetic association data to identify genetic risk variants that are then coupled with epidemiologic/environmental data. This strategy depends on the assumption that the genes influencing risk through interactions with environmental factors will also show significant independent/marginal association with disease. Thus, a 2-step approach can be taken in which stage 1 is a genome-wide scan of single nucleotide polymorphism (SNP) associations to identify a smaller set of SNPs passing a *P* value threshold to carry forward. In stage 2, the prioritized variant list is tested for SNP-environment interaction with the available environmental factors (8). This makes an important assumption about the relevance of the marginal SNP effect, but this strategy can be quite powerful when that assumption is met. Note that for quantitative outcomes, other criteria have been proposed to filter SNPs at step 1, including, for example, a test of heterogeneity of variance by genotypic classes (9).

Another approach to improving the power of interaction studies is to test for association with underlying quantitative traits for diseases that have a clear genetic component, referred to as endophenotypes. In some situations, such endophenotypes may be better measured and representative of the mechanistic pathway toward disease risk. For example, lipid profiles or glucose tolerance tests may be more appropriate for certain genetic association tests because they are closer to gene action than the disease outcome of heart disease or diabetes (10). Much like using marginal SNP association tests for selecting SNPs, by using endophenotypes that are closer to the molecular function of the gene, we will have more power to filter for potentially functional SNPs in subsequent gene-environment interaction (G×E) analyses. As we will discuss below, molecular endophenotypes such as gene expression may have sufficient power to filter environmental factors as well.

Finally, some studies have begun using prior biological knowledge to reduce the genome down to a set of candidate genes for G×E analyses. Here, there are a number of strategies for gene selection, each based on a set of assumptions:

- Evidence in the literature for variants in the gene region associated with the disease of interest
- Evidence in the literature of the gene associated with environmental factors of interest
- Genes related to the pathways where environmental factors may play a role
- Variants that are positioned in functional regions of the genome
- Use of public databases of G×E relationships and information about regulatory regions of the genome to filter the candidate genes and environmental factors

This is just a short list of options to reduce the genome to a smaller list of genes to test for G×Es. Such options are complemented by the use of model organism studies, which have several advantages for analyzing G×Es because environmental exposures can be carefully controlled and the genetic structure of the study can be leveraged to improve power (discussed further in McAllister et al. (11)). A variety of model organisms have been used for discovering G×Es (12) including yeast (13), *Drosophila* (14), and mouse (15, 16). Compared to human G×E studies, model organism studies allow the measurement of genetically similar individuals across multiple distinct environments. Considerable work has been done in this area, and it is outside the scope of this review. Thus, we focus our review on strategies that take advantage of molecular and cellular endophenotypes and multi-omics data.

USE OF -OMICS TO INFORM G×E DISCOVERY

Molecular and cellular endophenotypes provide a unique opportunity to identify genetic variants responsive to the environment at a more basic, mechanistic level. Unlike a GWAS, which may require tens of thousands of individuals to identify genetic variants associated with a phenotype, comparable studies of molecular and cellular endophenotypes using functional genomics have identified abundant associations with only dozens of individuals. This increased power to identify functional genetic loci has culminated in a wide diversity of quantitative trait studies for functional genomics data (or -omics), including epigenomes, methylomes, proteomes, and transcriptomes. Deciphering the role of these functional loci across human tissues has relied upon epigenome maps generated within large-scale projects where the data are publicly available, such as ENCODE (17) and the NIH Epigenomics Roadmap (18), in combination with expression quantitative trait locus (eQTL) studies from projects such as Multiple Tissue Human Expression Resource (MuTHER) (19) and Genotype-Tissue Expression (GTEx) (20).

Similar to using a marginal association test to prioritize important SNPs, epigenomic data (chromatin immunoprecipitation assays with sequencing (ChIP-seq), DNase I hypersensitive site sequencing (DNaseI-seq), and assay for transposase-accessible chromatin using sequencing (ATAC-seq)) can be used to identify enhancers and other regulatory elements in the genome and, subsequently, to prioritize variants positioned in those regions. Therefore, a straightforward application of -omics data is to identify responsive elements or variants at a molecular level in order to select candidate variants to test in G×E analyses.

In recent years, several studies have taken advantage of the increased power of molecular studies to map G×E effects. Barreiro et al. (21) mapped eQTLs in primary dendritic cells from 65 individuals before and after infection with *Mycobacterium tuberculosis* and identified 198 response eQTLs specific to either condition. This study demonstrated that nonnegligible numbers of G×E effects exist in the context of an infection and that mapping these variants and genes can be accomplished with limited numbers of individuals in well-controlled in vitro assays. Following this work, multiple studies have perturbed primary blood cells to elicit immune-response eQTLs (22–25). Beyond gene expression, response QTLs using chromatin accessibility assays have identified that immune-response eQTLs can be foreshadowed in the naïve state by regulatory variants influencing chromatin accessibility (26). This observation provides new opportunity to identify primed response variants that may underlie unexplained, noncoding, complex trait associations (27).

Response eQTLs need not be identified through in vitro perturbations alone; increasingly, studies of the interaction of observational variables such as age, sex, or behavior in cohorts with genetic and functional genomics data have identified G×E variants (28–31). For age- and sex-specific eQTLs, Yao et al. (28) focused on known complex disease-associated variants in a cohort of 5,254 individuals where whole blood gene expression was measured. They identified 10 age-specific and 14 sex-specific eQTLs, highlighting a notable scarcity of variants with strong effects given either variable. In addition to these variables, behaviors such as smoking, medication use, and exercise have been studied for G×E effects on gene expression; Knowles et al. (30) recently surveyed these variables using a novel approach for allele-specific expression to identify G×E effects for each environment. In both this study and Zhernakova et al. (31), investigators adopted the use of "proxy environments" to model unobserved perturbations, such as an individual's cell-type composition or infection status in discovery of interaction effects, providing a means to test previously unmeasured environments. Furthermore, the impact of genetics and in utero environments—including maternal smoking, birth weight, and birth order—have been observed to have large effects on an individual's methylome at birth (32, 33).

METHODOLOGICAL AND STUDY DESIGN ISSUES WITH THE USE OF -OMICS DATA

Use of -omics data in the discovery of G×Es presents unique methodological issues. First, there is the issue of tissue specificity. The identity of the most effective tissues for interaction testing is not always clear. For instance, in pursuing genetic contributors to adverse drug reactions, we could analyze either drug-metabolizing tissues such as the liver or the tissues where adverse effects are presented. Obtaining these tissues can be quite challenging, and even if samples can be obtained, we must consider whether such effects can be more easily and less invasively discovered from accessible patient tissues such as skin or blood. Indeed, in the study of adverse drug reactions to statins, the link to the *GATM* gene was observed through analysis of statin-response eQTLs in lymphoblastoid cells rather than muscle or liver (34). Determining the extent of tissue specificity of genetic effects is still a major challenge. Studies such as the Genotype-Tissue Expression project have begun to comprehensively identify tissue-shared and tissue-specific effects (20). Designing G×E studies in the correct tissues will further benefit from increasingly rich epigenomic maps, such as ENCODE and the NIH Epigenomics Roadmap. By identifying the regions that are differentially accessible in response to an environmental perturbation in a few tissues, one can use these large maps to gain insight into the relative benefits of G×E testing in different tissues.

The problem of statistical power and adequate sample size can significantly impede G×E studies (35). To study the effect of a cellular perturbation in vitro as in the study by Fairfax et al. (22), sample sizes of a few hundred individuals per condition or a few time points may be required. The study design should include -omics analysis (RNA sequencing, DNase-Seq, ATAC-Seq, etc.) at every condition or time point. If measurements are made on the same individuals, then statistical methods need to account for within-person correlation, an adjustment that may decrease power. This design is feasible for in vitro studies. Only recently have in vivo studies begun to track large numbers of individuals with repeated -omics measurements. These emerging studies pose additional challenges. In vivo studies need to account for the potential impact of multiple environments, increased variability in sample collection and measures of the environment, and increased challenges in causal inference.

Other methodological challenges facing G×E studies arise largely from the inherent properties of -omics data. Data generated from a specific -omics technology are modeled better by certain distributions than others. Knowledge of the underlying distribution of the data governs the models chosen to detect G×E effects. For example, it has been widely shown that gene expression as measured by RNA sequencing can be adequately modeled by a negative binomial distribution (36). For gene expression measured using microarrays, one might chose a linear model with an interaction term (gene-expression variable × environment). However, for RNA-sequencing data, a generalized linear model accounting for overdispersion in the count data may be preferable. Another consideration is that -omics data encounters challenges from known and unknown confounders. Known confounders may include sex, age, or batch. In fact, known biological confounders may appear as environmental effects that we wish to test in a subset of our analysis. Many methods have been developed to remove these unwanted effects in the context of eQTL studies from simple linear regression and/or principal components analysis to more advanced techniques (probabilistic estimation of expression residuals (37), HCP (38), svaseq (39)). For in vitro studies, these methods are typically applied to each condition or time point separately. For in vivo studies, one must take care not to remove the environmental effect to be tested. Given the diversity of cellular perturbations that can be assayed and studied for G×E effects, an ongoing bottleneck is selecting specific environments to assay. Moyerbrailean et al. (40) demonstrated an effect pipeline for assessing G×E effects across 250 environments (50 perturbations in 5 cell types) where all perturbations were tested using low-pass transcriptome sequencing. Only those perturbations with significant differences were then carried forward for deeper sequencing and assessment of context-specific, allele-specific expression. Using this approach, they identified 215 genes with G×Es, of which nearly 50% were implicated in previous

GWASs, thereby providing a rich resource of candidate hypotheses for further investigation.

Ultimately, estimating the degree to which genetics and environment contribute to molecular and cellular phenotypes poses a principal challenge to G×E study design. Multiple gene-expression studies have used family relationships to determine the relative contributions of genetics and environment, reporting average heritability from 0.1 to 0.26 (41, 42). The majority of studies calculate heritability from total expression levels, but allele-specific expression is increasingly complementing these estimates. Buil et al. (41) used allele-specific expression measured in a twin study of approximately 400 female twin pairs with RNA- sequencing data from different tissues (fat, lymphoblastoid cell lines, skin, and blood) to measure the relative contribution of genetics and environment. In support of previous studies, they found little evidence for shared environmental effects. However, they found significant evidence of unique or individual-specific environmental effects, explaining approximately 10%–20% of expression variation in each of the tissues, on par with cis or cis-trans effects, and they further estimated that 38%–49% of variance observed in allele-specific expression was not explained by additive effects and was due to gene-gene or G×E effects. Despite this, limitations on power influenced the discovery of specific G×E associations, highlighting that either larger studies or candidate studies were required.

INTEGRATIVE AND PATHWAY-BASED STRATEGIES

Genes do not act in isolation; they function through physical, metabolic, and chemical reactions with other genes in large pathways and networks. Yet when we look for associations between genes and disease outcomes or even G×Es, we tend to treat each gene independently. If we embrace the complexity and relationships between genes in pathways, we may increase our power to detect and interpret our findings. Many researchers describe pathway approaches as being antithetical to the unbiased genome-wide perspective. Thomas (43) describes ways that these powerful pathway-based approaches can further enhance power and insights rather than replace the GWAS approach. While taking an agnostic GWAS approach to generate the genetic data might be beneficial to ensure that all regions of the genome are explored, these data can be married with hierarchical modeling strategies that exploit pathway knowledge when we perform analyses of genome-wide data (43).

Exploring candidate pathways and using biological knowledge related to the environment in the gene selection process can potentially be a powerful alternative to the current paradigm. Two strategies can be applied. First, as for gene-based and pathway-based analyses of marginal SNP effects, we can perform an agnostic search for enrichment of interaction effects. Indeed, existing methods such as gene-set enrichment analysis (44) only assume genome-wide *P* values of the test considered (marginal effect, a priori) following a uniform distribution under the null hypothesis of no enrichment. Therefore, these methods can be directly applied to the standard 1-degree-of-freedom test of interaction performed on a genome-wide scale. Because it relies on established methodologies, some groups have already started to apply such strategies. For example, Wei et al. (45) combined gene-based testing with pathway enrichment analyses to look for G×E associations with lung-cancer susceptibility.

The second potential strategy is to use pathway information to reduce the search space for interactions, as in candidate-based approaches. To facilitate selection of candidate variants and study of G×Es, the CardioGxE database has extensively curated G×E variants in the literature (46). Further, as discussed in previous sections, a common strategy in G×E screening is to assume SNPs involved in interaction effects also display a marginal effect. Similarly, one can argue that SNPs involved in interactions might be enriched in pathways related to the outcome or the exposure in question. Building on this idea, Rava et al. (47) proposed a strategy to select genes for G×E for associations with asthma (a trait for which earlier G×E studies did not find strong evidence of associations). They selected the canonical pathways that the set of candidate genes belong to and included biological knowledge related to the environment in their gene selection process. Their approach reduced the gene list to a small and focused set for G×E analysis (47). Similar approaches have been adopted for other diseases as well. Huang and Hu (48) used these same strategies to look for G×Es and associations with obesity. Tang et al. (49) performed gene-based gene × smoking interaction analysis, followed by enrichment analysis of nominally interacting genes in canonical biological pathways, and found that the axonal guidance signaling pathway interacted with smoking to modify the risk of pancreatic cancer. Of note, exactly the same pathway was identified to be enriched with recurrent somatic point mutations and copy number aberrations by an exome-sequencing study of pancreatic tumors (50). This study also suggested a negative correlation between somatic mutations in the axonal guidance pathway and smoking—nonsmokers were more likely to have somatic mutations in this pathway than smokers. For the first time, both GWAS and somatic mutation data pointed to the same biological pathway that might interact with smoking in modifying cancer risk.

To integrate this type of biological knowledge into an analysis, bioinformatics tools such as Biofilter can be used. Biofilter is a knowledge integration tool developed to allow for annotation, filtering, and model building of genomic data (51, 52). The underlying database of biological knowledge within Biofilter is called the Library of Knowledge Integration (LOKI) and consists of multiple public database sources such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/), Gene Ontology (http://www.geneontology.org/), Pfam (http://pfam.xfam.org/), and the GWAS Catalog (https://www.ebi.ac.uk/gwas/). These sources are linked in the Library of Knowledge Integration so that a user can provide a list of either SNPs or genes and get the SNPs annotated by gene, the genes annotated by group (pathway, protein family, etc.), and the lists of genes that belong to specific groups. These annotations can be used to filter gene sets into subsets prior to association analyses. For example, in cataract susceptibility, Hall et al. (53) used Biofilter to identify potential gene-gene interactions in the eMERGE network. This same process could be used for G×E analysis with a candidate gene list based on either the disease or the environmental-factor candidates; the user would provide the candidate gene list, and Biofilter would provide all of the other genes that are linked to the genes in the user's list based on the Library of Knowledge Integration. Tools like Biofilter make performing candidate-pathway, or gene-set, approaches more efficient because they integrate knowledge from multiple public data sources and enable researchers to use all of the sources simultaneously.

Because of their novelty and the limited number of applications that have been developed so far, it is challenging to assess the relevance of these approaches. Obviously, their performances rely on the validity of the underlying statistical and biological assumptions, such as valid and correct biological/pathway knowledge. For example, in the later candidate-based approach, substantial gain in power might be achieved if G×E are indeed enriched in outcome/exposure pathways, but might have decreased power otherwise. Also, as in the case of marginal-effects testing of genes and pathways, the potential success of these strategies lies in their ability to cope with multiple causal genes whose effects might be heterogeneous across populations. Focusing on single-variant association signals, as in standard G×E GWAS screening, implicitly assumes the effects of each variant is large enough and homogeneous enough that it can be replicated across populations. However, when effects are small, and variability due to other uncontrolled risk factors is high, gene- and pathway-based tests, which aggregate information, might be more efficient than those using SNPs as the testing unit. Thus, it is anticipated that these pathway-based approaches could increase power by magnitudes, assuming that the knowledge used is accurate. This is a critical assumption. Direct comparisons of the statistical power of the different approaches are not possible at this time. However, as more applications of these approaches are published, we will learn more about the gain in power through the use of these pathway-based strategies.

FUTURE DIRECTIONS AND CONCLUSIONS

We have described several different biological considerations that can be used to improve power and increase flexibility in using prior knowledge and/or multiple data types in G×E analyses. The wealth of biological knowledge that has been discovered over the past two decades is astonishing. It is clearly to our benefit to include this knowledge; this is particularly important because it can help us find biologically meaningful G×Es. Also, this knowledge may be useful to help identify interactions with rare or low-frequency variants that may have been missed previously.

An additional important consideration is that of replication of the G×E effects. In genetics, replication of association has become the gold standard (54). Here, to avoid an inflation of false positives, the field looks for replication of the precise result in multiple, independent data sets. This is an important strategy under an assumption where we believe that one particular SNP and one particular exposure measurement are important for the disease of interest. However, if we start to consider gene-based models or pathway models, what is the "model" that we need to replicate? If we see several genes from a particular pathway associating in a G×E with a particular trait of interest in one data set and then 2 different genes from the same pathway associating in a G×E in a second data set, is that replication? This would not be a replication under the traditional definition (the same independent variables combined in the same statistical model with the same direction of effect). We need to consider this apparent contradiction more carefully as we expand analyses to accommodate gene-based or pathway-based approaches.

Our ability to integrate information from multi-omics approaches, the literature, and other public knowledge sources provides tremendous power to deal with the challenges inherent in big-data analyses. In the coming years, we expect to see more G×E analyses incorporating biological knowledge–driven strategies similar to those described here. These approaches will continue to evolve and expand as we learn more about the relationships between the genome and the exposome in the architecture of complex traits.

ACKNOWLEDGMENTS

Author affiliations: Department of Biochemistry and Molecular Biology, Center for Systems Genomics, Pennsylvania State University, University Park, Pennsylvania (Marylyn D. Ritchie); Biomedical and Translational Informatics, Geisinger Health System, Danville, Pennsylvania (Marylyn D. Ritchie); Departments of Genetics, Stanford University School of Medicine, Stanford, California (Joe R. Davis III); Department of Epidemiology, Harvard T.H. School of Public Health, Boston, Massachusetts (Hugues Aschard, Peter Kraft); Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France (Hugues Aschard); Department of Computer Science, Johns Hopkins University, Baltimore, Maryland (Alexis Battle); Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California (David Conti, Duncan C. Thomas); Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York (Mengmeng Du); Department of Computer Science, University of California, Los Angeles, Los Angeles, California (Eleazar Eskin); Department of Human Genetics, Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California (Eleazar Eskin); Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland (M. Daniele Fallin); Biostatistics and Biomathematics Program, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington (Li Hsu); Division of Informatics, Department of Biostatistics and Epidemiology, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania (Jason H. Moore); Department of Public Health Sciences, University of Chicago Medicine and Biological Sciences, Chicago, Illinois (Brandon L. Pierce); Department of Human Genetics, University of Chicago Medicine and Biological Sciences, Chicago, Illinois (Brandon L. Pierce); Comprehensive Cancer Center, University of Chicago Medicine and Biological Sciences, Chicago, Illinois (Brandon L. Pierce); Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington (Stephanie A. Bien); Department of Biostatistics, University of Texas MD Anderson Cancer

Center, Houston, Texas (Peng Wei); Department of Genetics, Stanford University School of Medicine, Stanford, California (Stephen B. Montgomery); and Department of Pathology, Stanford University School of Medicine, Stanford, California (Stephen B. Montgomery).

M.D.R. and S.B.M. contributed equally to this work.

Research reported in this publication was supported by National Institute of Environmental Health Sciences; National Library of Medicine; National Institute of Allergy and Infectious Diseases; National Cancer Institute; National Heart, Lung and Blood Institute; National Human Genome Research Institute; and National Institute of Mental Health of the National Institutes of Health under award numbers R21HG007687 to H.A.; R01CA140561, R01CA201407, P01CA196569 to D.C.; P30CA008748, R25CA094880 to M.D.; R01CA189532, R01CA195789, and P01CA53996 to L.H.; R01LM010098, R01AI116794, and P30ES013508 to J.H.M.; R01ES023834, R21ES024834 to B.L.P.; R01CA169122, R01HL116720 and R21HL126032 to P.W. E.E. is supported by R01ES022282. S.B.M. is supported by the National Institutes of Health through R01HG008150, R01MH101814, U01HG007436, and U01HG009080. J.R. D. is supported by a Lucille P. Markey Biomedical Research Stanford Graduate Fellowship. J.R.D. acknowledges the Stanford Genome Training Program (SGTP; NIH/NHGRI T32HG000044). This work is funded, in part, under a grant with the Pennsylvania Department of Health (SAP 4100070267) to M.D.R.

The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest: none declared.

REFERENCES

1. Hindorff LA, MacArthur J, Morales J, et al. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies. Accessed April 1, 2013.
2. Maher B. Personal genomes: the case of the missing heritability. *Nature*. 2008;456(7218):18–21.
3. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747–753.
4. Pekkanen J, Pearce N. Environmental epidemiology: challenges and opportunities. *Environ Health Perspect*. 2001; 109(1):1–5.
5. Patel CJ, Ioannidis JP. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *J Epidemiol Community Health*. 2014;68(11):1096–1100.
6. Zappa C, Mousa SA. Non-small cell lung cancer: current treatment and future advances. *Transl Lung Cancer Res*. 2016; 5(3):288–300.
7. Engel C, Fischer C. Breast cancer risks and risk prediction models. *Breast Care (Basel)*. 2015;10(1):7–12.
8. Dai JY, Kooperberg C, Leblanc M, et al. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*. 2012;99(4): 929–944.
9. Paré G, Cook NR, Ridker PM, et al. 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*. 2010;6(6):e1000981.
10. Wright A, Burden AC, Paisey RB, et al. Sulfonylurea inadequacy: efficacy of addition of insulin over 6 years in patients with type 2 diabetes in the UK Prospective Diabetes Study (UKPDS 57). *Diabetes Care*. 2002;25(2):330–336.
11. McAllister K, Mechanic LE, Amos C, et al. Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *Am J Epidemiol*. 2017;186(7):753–761.
12. Flint J, Mackay TF. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res*. 2009;19(5):723–733.
13. Smith EN, Kruglyak L. Gene-environment interaction in yeast gene expression. *PLoS Biol*. 2008;6(4):e83.
14. Zhou S, Campbell TG, Stone EA, et al. Phenotypic plasticity of the Drosophila transcriptome. *PLoS Genet*. 2012;8(3):e1002593.
15. Valdar W, Solberg LC, Gauguier D, et al. Genetic and environmental effects on complex traits in mice. *Genetics*. 2006;174(2):959–984.
16. Parks BW, Nam E, Org E, et al. Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell Metab*. 2013;17(1):141–152.
17. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489(7414):57–74.
18. Roadmap Epigenomics Consortium, A Kundaje, W Meuleman, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317–330.
19. Nica AC, Parts L, Glass D, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet*. 2011;7(2):e1002003.
20. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648–660.
21. Barreiro LB, Tailleux L, Pai AA, et al. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc Natl Acad Sci USA*. 2012;109(4):1204–1209.
22. Fairfax BP, Humburg P, Makino S, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*. 2014;343(6175):1246949.
23. Lee MN, Ye C, Villani AC, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*. 2014;343(6175):1246980.
24. Ye CJ, Feng T, Kwon HK, et al. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science*. 2014;345(6202):1254665.
25. Çalışkan M, Baker SW, Gilad Y, et al. Host genetic variation influences gene expression response to rhinovirus infection. *PLoS Genet*. 2015;11(4):e1005111.
26. Alasoo K, Rodrigues J, Mukhopadhyay S, et al. Genetic effects on chromatin accessibility foreshadow gene expression changes in macrophage immune response. *bioRxiv*. (doi:10.1101/102392).
27. Chun S, Casparino A, Patsopoulos NA, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat Genet*. 2017;49(4):600–605.
28. Yao C, Joehanes R, Johnson AD, et al. Sex- and age-interacting eQTLs in human complex diseases. *Hum Mol Genet*. 2014;23(7):1947–1956.
29. Kukurba KR, Parsana P, Balliu B, et al. Impact of the X chromosome and sex on regulatory variation. *Genome Res*. 2016;26(6):768–777.

30. Knowles DA, Davis JR, Edgington H, et al. Allele-specific expression reveals interactions between genetic variation and environment. *Nat Methods*. 2017;14(7):699–702.
31. Zhernakova DV, Deelen P, Vermaat M, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet*. 2017;49(1):139–145.
32. Teh AL, Pan H, Chen L, et al. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome Res*. 2014;24(7):1064–1074.
33. Galanter JM, Gignoux CR, Oh SS, et al. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *Elife*. 2017;6:e20532.
34. Mangravite LM, Engelhardt BE, Medina MW, et al. A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature*. 2013;502(7471):377–380.
35. Aschard H. A perspective on interaction effects in genetic association studies. *Genet Epidemiol*. 2016;40(8):678–688.
36. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
37. Stegle O, Parts L, Piipari M, et al. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*. 2012;7(3):500–507.
38. Mostafavi S, Battle A, Zhu X, et al. Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One*. 2013;8(7):e68141.
39. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res*. 2014;42(21):e161.
40. Moyerbrailean GA, Richards AL, Kurtz D, et al. High-throughput allele-specific expression across 250 environmental conditions. *Genome Res*. 2016;26(12):1627–1638.
41. Buil A, Brown AA, Lappalainen T, et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet*. 2015;47(1):88–91.
42. Wright FA, Sullivan PF, Brooks AI, et al. Heritability and genomics of gene expression in peripheral blood. *Nat Genet*. 2014;46(5):430–437.
43. Thomas D. Gene–environment-wide association studies: emerging approaches. *Nat Rev Genet*. 2010;11(4):259–272.
44. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102(43):15545–15550.
45. Wei S, Wang LE, McHugh MK, et al. Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility. *Carcinogenesis*. 2012;33(8):1531–1537.
46. Parnell LD, Blokker BA, Dashti HS, et al. CardioGxE, a catalog of gene-environment interactions for cardiometabolic traits. *BioData Min*. 2014;7:21.
47. Rava M, Ahmed I, Demenais F, et al. Selection of genes for gene-environment interaction studies: a candidate pathway-based strategy using asthma as an example. *Environ Health*. 2013;12:56.
48. Huang T, Hu FB. Gene-environment interactions and obesity: recent developments and future directions. *BMC Med Genomics*. 2015;8(suppl 1):S2.
49. Tang H, Wei P, Duell EJ, et al. Axonal guidance signaling pathway interacting with smoking in modifying the risk of pancreatic cancer: a gene- and pathway-based interaction analysis of GWAS data. *Carcinogenesis*. 2014;35(5):1039–1045.
50. Biankin AV, Waddell N, Kassahn KS, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*. 2012;491(7424):399–405.
51. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput*. 2009;368–379.
52. Pendergrass SA, Frase A, Wallace J, et al. Genomic analyses with Biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData Min*. 2013;6(1):25.
53. Hall MA, Verma SS, Wallace J, et al. Biology-driven gene-gene interaction analysis of age-related cataract in the eMERGE network. *Genet Epidemiol*. 2015;39(5):376–384.
54. NCI-NHGRI Working Group on Replication in Association Studies, Chanock SJ, Manolio T, et al. Replicating genotype-phenotype associations. *Nature*. 2007;447(7145):655–660.