Big Data
Volume 3 Number 3, 2015
DOI: 10.1089/big.2015.0013

ORIGINAL ARTICLE

How Genetic and Other Biological Factors Interact with Smoking Decisions

Laura Bierut,[1,*] and David Cesarini[2]

Abstract

Despite clear links between genes and smoking, effective public policy requires far richer measurement of the feedback between biological, behavioral, and environmental factors. The Kavli HUMAN Project (KHP) plans to exploit the plummeting costs of data gathering and to make creative use of new technologies to construct a longitudinal panel data set that would compare favorably to existing longitudinal surveys, both in terms of the richness of the behavioral measures and the cost-effectiveness of the data collection. By developing a more comprehensive approach to characterizing behavior than traditional methods, KHP will allow researchers to paint a much richer picture of an individual's life-cycle trajectory of smoking, alcohol, and drug use, and interactions with other choices and environmental factors. The longitudinal nature of KHP will be particularly valuable in light of the increasing evidence for how smoking behavior affects physiology and health. The KHP could have a transformative impact on the understanding of the biology of addictive behaviors such as smoking, and of a rich range of prevention and amelioration policies.

Key words: smoking; genetics; deep phenotyping; smoking cessation

Introduction

Despite the fact that it has long been understood that smoking is a leading modifiable risk factor for poor health,[1] estimates suggest that tobacco use continues to be responsible for nearly one in five U.S. deaths.[2] Even though the development of smoking cessation and prevention strategies has been a major priority for policy makers for quite some time, progress has been hampered by our as-of-yet imperfect understanding of the complex genetic and environmental etiology of smoking behavior. In an era of rapid technological advances in the measurement and analysis of DNA, the understanding of robustly established—but difficult-to-interpret—genetic associations with smoking behavior made possible by "big data" can be substantially enhanced through careful follow-up analyses in rich longitudinal panels with data of high quality.

The advent of genome-wide association studies (GWAS) has massively increased the ability to identify genes that impact deleterious behaviors. In particular, robust and biologically plausible associations have been discovered between smoking and genes. Yet, different facets of smoking behavior—initiation, intensity, and cessation—have distinct biologic and environmental contributors. To test hypotheses about genetic effects on smoking, it is therefore critical to have reliable measures of the various facets of smoking behavior over the life cycle. Yet, current behavioral measures, such as maximum level of smoking at any point in the life cycle, remain crude.

By radically improving measurement of behavioral phenotypes, the Kavli HUMAN Project (KHP) will clarify links between biology and health-impacting behaviors such as smoking. For example, self-reported smoking quantities can be cross-checked against credit-card records on cigarette purchases and supplemented with information from medical records about health conditions associated with tobacco use. Direct biological measurement of smoking markers, such as cotinine—a compound formed after nicotine enters the body—and exhaled carbon monoxide—a measure

[1]*Washington University in St. Louis, St. Louis, Missouri.*
[2]*New York University, New York, New York.*

*Address correspondence to: *Laura Bierut, Washington University in St. Louis, 660 South Euclid, St. Louis, MO 63110*, E-mail: bierutl@psychiatry.wustl.edu

of exposure to smoked combustible cigarettes—will also be informative.

The KHP will particularly enrich the understanding of feedback mechanisms between biology and behavior. For example, studies have begun to identify several genes whose levels of methylation are associated with smoking behavior. Whether these changes can help explain some of the biological pathways through which smoking ultimately impacts lung health and lung cancer is a vibrant area of research. Longitudinal data sets with rich behavioral and biological measures can be an invaluable resource for enhancing the understanding of the links between smoking and health.

Genetics of Smoking

Beginning around 2005, medical genetics research began to undergo a paradigm shift, moving to GWAS. In these studies, made feasible by technological advances, researchers test the outcome of interest for association with each of the measured single-nucleotide polymorphisms (SNPs). Because of the large number of hypotheses tested in a GWAS, a SNP association is considered to be established only if it reaches the "genome-wide significance" threshold of $p < 0.00000005$. Adequate statistical power at this stringent significance threshold requires very large samples. Since individual samples are generally too small, many GWAS are conducted within research consortia that meta-analyze results from multiple samples and countries.

Empirically, it is now well established that results from such GWAS replicate very consistently.[3] There are several reasons for the robustness of GWAS findings (see Rietveld et al.[4] for a discussion). Before the modern era of GWAS, most molecular genetic studies of smoking had been candidate gene studies, which focused exclusively on studied variations in genes in biological systems known to play an important role in nicotine addiction. The replication record of these early studies turned out to be disappointing, and the estimates of the effect sizes were often highly heterogeneous across studies.[5] An influential review[6] concluded that the "evidence for a contribution of specific genes to smoking behavior remains modest." Ten years later, the GWAS have uncovered a handful of genetic associations with smoking behavior for which the evidence is very strong and the replication record is excellent.

A landmark event in the study of the genetics of smoking was the publication of the first GWAS of smoking in *Nature*,[7] along with two studies of lung cancer in *Nature*[8] and *Nature Genetics*.[9] This work was followed by three GWAS of smoking behavior in the May 2010 issue of *Nature Genetics*.[10–12] By far the strongest results came from a set of SNPs located in the chromosome 15 cluster of virtually adjacent nicotinic receptor genes (*CHRNA3*, *CHRNA5*, and *CHRNB4*), which were identified in all studies as a risk factor for heaviness of smoking defined by number of cigarettes smoked per day (CPD), as well as the strongest genetic risk for the development of lung cancer. The SNP rs16969968, known colloquially among researchers as "Mr. Big," is widely believed to be the causal variant underlying the signal. In particular, it is known to cause an amino acid change in the alpha-5 subunit of the nicotinic receptors, and experiments have found that this change alters the responsiveness of the nicotinic receptors to nicotine.[13]

Despite many strengths, the GWAS also have some obvious limitations. First, it is often necessary to sacrifice phenotype quality to attain sample sizes needed for studies to have adequate power to detect associations. As a result, it is not always easy to interpret an observed association. For example, the "TAG" study[11] combined quite different measures in a single study: some cohorts asked smokers about their maximum daily consumption at peak consumption, whereas others asked about contemporaneous consumption. Moreover, GWAS are useful for identifying genetic signals, but are of limited value for understanding how an identified genetic effect might vary across environmental conditions. In the case of smoking, it is *a priori* plausible that such interactions are often of first-order importance.

Thus, GWAS are of great value for detecting real and replicable genetic associations, but they are merely a necessary first step toward the more ambitious twin goals of identifying the ensemble of genes that, along with environmental factors, account for heterogeneity across individuals, and understanding how environmental factors can amplify or dampen genetic risk. Credibly establishing such pathways requires rich longitudinal measures of behavior, biological markers, and environmental factors. Because no such data set presently exists, the KHP could potentially fill an important void.

Figure 1 shows why it is likely that this void will continue to grow in the coming years, as larger and larger discovery samples lead to the discovery of more and more genetic associations with various complex outcomes. For example, the first study of schizophrenia identified a single polymorphism,[14] but the availability of larger and larger samples has brought the number up to 108.[15] Early studies of height identified 10–20

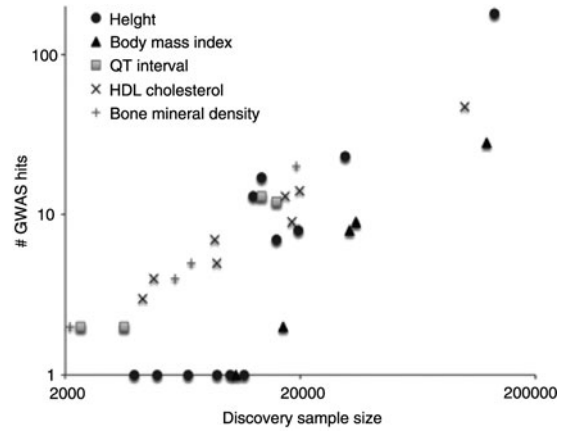

FIG. 1. Number of polymorphisms identified as a function of meta-genome-wide association studies sample size. Note the logarithmic scale on the *x*-axis. (Source: Visscher et al.[3]).

polymorphisms,[16–18] whereas new research by the GIANT consortium,[17] based on a sample of 250,000 individuals, identifies 700.

It seems exceedingly likely that in the coming years, we will similarly be awash in genetic associations with smoking phenotypes as well as measures of other substance use. To have the greatest scientific impact, these associations will require interpretation and follow-up work on behavioral and biological mechanisms. Below, three concrete examples are given of the complementarities believed to exist between the use of "big data" for gene discovery and the use of high-quality longitudinal data with rich behavioral, biological, and environmental measure to refine the understanding of mechanisms.

Improving Phenotype Measurement by Leveraging Multiple Data Sources

A major challenge in addiction research is phenotype measurement. By combining conventional longitudinal survey-based measures with novel ways of measuring smoking behavior, the KHP will allow researchers to paint a much richer picture of an individual's life-cycle trajectory of smoking, alcohol, and other substance use. For example, measuring substance use behavior is associated with three major difficulties. First, subjects who are surveyed on a single occasion may exhibit recall biases. Second, because of the social stigma associated with substance use, some respondents may systematically color their responses. Third, many surveys ask about substance use at a single point in time, and responses to these questions may be poor proxies for an individual's life cycle of substance use behavior.

The KHP data could be leveraged in a number of ways to obtain more reliable measures of substance use. By tracking people longitudinally, it is possible to measure changes in substance use behavior much more accurately over time. Moreover, a number of other data sources could be used to improve phenotype measurement and validate survey responses. For example, self-reported smoking quantities could be cross-checked against credit-card records on cigarette purchases. There are also a number of well-known biomarkers for smoking behaviors and other substance use. Cotinine, which can be measured from saliva,[19] is often used to obtain an objective measure of an individual's exposure to tobacco.[20] An exciting development in recent years is the fact that it is becoming feasible to measure DNA methylation, an epigenetic mechanism for the regulation of gene expression. Epigenome-wide association studies have identified several genes whose methylation is strongly associated with smoking behavior.[21] Finally, survey questions could be supplemented with information from medical records about health conditions associated with tobacco use (such as diagnostic codes for pulmonary disease and lung cancer) or diagnostic codes for treatment of tobacco use and dependence.

Existing genetic studies suggest that the genetic architecture of different facets of smoking behavior—initiation, intensity, cessation—show quite modest genetic overlap. To test hypotheses about genetic effects on smoking, it is therefore critical to have reliable measures of the various facets of smoking behavior over the life cycle.

Illuminating Biological Consequences of Health Behaviors

A very robust finding emerging from the epigenome-wide association studies of methylation conducted to date is that smoking is associated with the methylation of many genes. Whether these methylation patterns can help to explain some of the biological pathways through which smoking ultimately impacts lung health[22] and lung cancer[23] is a vibrant area of research. The KHP would be a valuable resource for testing hypotheses about several of the genes whose methylation is believed to play an important role in the causal pathways from smoking to poor health. Most studies measure methylation from the blood, but methylation can also be measured in other types of tissue, including saliva, which is easier and cheaper to collect on a large scale.

Gene–Environment Interactions and Behavioral Pathways

Finally, the KHP data could be a valuable resource for testing hypotheses about gene–environment (G×E) interactions. Efforts to understand interactions between environmental factors and tobacco and alcohol consumption are already well underway.[24,25] A major challenge for studies of G×E is that the measures of environmental exposures are often imperfect; the KHP's ambitious plans for gathering high-quality data on life events and other environmental variables would thus fill an important void. A second challenge is that to deliver convincing answers, G×E studies need to have adequate statistical power.[26] The large and richly phenotyped KHP sample would thus help to overcome two serious obstacles to scientific progress in this area. Indeed, the large sample would permit meaningful analyses even in fairly narrowly defined subgroups. Hypotheses about interactions could be tested in suitably selected subsamples through randomized interventions.

In studies of G×E, it is also envisioned that there will be large gains from collaborations between geneticists, who contribute critical biological expertise, and economists, who are well trained in teasing out causal relationship from observational data. In the social sciences, controlled experiments are not always a feasible research strategy for establishing causality. Confronted with this reality, researchers have shown great ingenuity in developing methods to tease out causal relationships from "quasi-experiments," events that produce variation that plausibly resembles the experimental variation generated by a controlled experiment (for an overview, see Angrist and Pischke[27]). For example, studies have studied lottery winners to study the causal impact of wealth on labor supply,[28] and adoptees assigned to families using plausibly random mechanisms to learn about the impact of family environment on child outcomes.[29] During the course of the study, it is likely that some subjects will be exposed to plausibly exogenous environmental insults, for example a large unanticipated bequest or serious injury from an accident. Such naturally occurring variation can be leveraged to gain insight into causal interactions between genetic and environmental factors.

Implementation in the KHP

Investigation of factors that contribute to smoking decisions would utilize the following KHP data sets, among others: (a) Smoking use data would be available via medical history and records forward, and through KHP's biological samples (see below), as well as by mining for purchase of tobacco products in the financial data. Mining financial data would offer additional benefits over the limitations of survey-only methods due to its continuous basis, and it would also help confirm actual cessation of smoking versus "claimed" cessation. (b) Air quality and ambient noise levels would be measured via sensors placed in the home. (c) Exposure to toxins and other chemicals would be measured via silicone wristbands worn periodically. (d) Information on financial status and participation in government assistance programs (Supplemental Nutrition Assistance Program, Social Security, Temporary Cash Assistance to Needy Families) would be available via financial data gathered using a combination of automated and survey-based methodologies.

The impact of smoking decisions on health would be analyzed via the following KHP data sets, amongst others: (a) Medical information on study participants' health would be available from the medical history and records going forward (medical records, doctors' notes, hospital records, dental records). Prescription data would be gathered via the NY State Prescription database. This information would be complemented by the Statewide Planning and Research Cooperative System database and KHP's own tests: blood tests (blood metals, vitamins, lipids, glucose and other biomarkers), urine and hair tests (smoking, alcohol and substance use) every 3 years. (b) Information of a genetic nature, including telomere length, would be gathered via whole genome sequencing of blood samples for adults (saliva for children) performed at study intake. In addition, data on variation in epigenetics would be gathered via triennially performed assays.

Conclusion

It has been emphasized that there is no conflict between research approaches that leverage enormous data sets to discover basic patterns of association and research approaches leveraging rich longitudinal data sets to test specific causal hypotheses. Rather, two approaches should be viewed as mutually reinforcing and necessary for making progress on designing effective health interventions.

Author Disclosure Statement

Laura J. Bierut is listed as an inventor on Issued U.S. Patent 8,080,371, "Markers for Addiction" covering the use of certain SNPs in determining the diagnosis,

prognosis, and treatment of addiction. For David Cesarini, no competing financial interests exist.

References

1. Centers for Disease Control and Prevention. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. Atlanta, Georgia 2010. Available online at www.cdc.gov/tobacco/data_statistics/sgr/2010/index.htm?s_cid=cs_1843 (last accessed June 1, 2015).
2. Mokdad AH, Marks JS, Stroup DF, et al. Actual causes of death in the United States, 2000. JAMA. 2004;291:1238–1245.
3. Visscher PM, Brown MA, McCarthy MI, et al. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.
4. Rietveld CA, Conley D, Eriksson N, et al. Replicability and robustness of GWAS for behavioral traits. Psychol Sci. 2014;25:1975–1986.
5. Aljasir B, Ioannidis JP, Yurkiewich A, et al. Assessment of systematic effects of methodological characteristics on candidate genetic associations. Hum Genet. 2013;132:167–178.
6. Munafò M, Clar T, Johnstone E, et al. The genetic basis for smoking behavior: a systematic review and meta-analysis. Nicotine Tob Res. 2004;6:583–597.
7. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452:638–642.
8. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 2008;452:633–637.
9. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat. Genet. 2008;40:616–622.
10. Liu JZ, Tozzi F, Waterworth DM, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. 2010;42:436–440.
11. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010;42:441–447.
12. Thorgeirsson TE, Gudbjartsson DF, Surakka I, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet. 2010;42:448–453.
13. Bierut LJ, Stitzel JA, Wang JC, et al. Variants in nicotinic receptors and risk for nicotine dependence. Am J Psychiatry 2008;165:1163–1171.
14. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;460:748–752.
15. Ripke S, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature 2014;511:421–427.
16. Gudbjartsson DF, Walters GB, Thorleifsson G, et al. Many sequence variants affecting diversity of adult human height. Nat Genet. 2008;40:609–615.
17. Lango Allen H, Estrada K, Lettre G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 2010;467: 832–838.
18. Weedon MN, Lango H, Lindgren CM, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet. 2008;40:575–583.
19. Etter JF, Vu Duc T, Perneger TV. Saliva cotinine levels in smokers and nonsmokers. Am J Epidemiol. 2000;151:251–258.
20. Benowitz NL. Cotinine as a biomarker of environmental tobacco smoke exposure. Epidemiol Rev. 1996;18:188–204.
21. Lee KWK, Pausova Z. Cigarette smoking and DNA methylation. Front Genet 2013;4:132.
22. Zong DD, Ouyang RY, Chen P. Epigenetic mechanisms in chronic obstructive pulmonary disease. Eur Rev Med Pharmacol Sci. 2015;19:844–856.
23. Huang T, Chen X, Hong Q, et al. Meta-analyses of gene methylation and smoking behavior in non-small cell lung cancer patients. Sci Rep. 2015;5 8897.
24. Grucza RA, Johnson EO, Krueger RF, et al. Incorporating age at onset of smoking into genetic models for nicotine dependence: evidence for interaction with multiple genes. Addict Biol. 2010;15:346–357.
25. Olfson E, Edenberg HJ, Nurnberger J Jr, et al. An ADH1B variant and peer drinking in progression to adolescent drinking milestones: evidence of a gene-by-environment interaction. Alcohol Clin Exp Res. 2014;38: 2541–2549.
26. Duncan LE, Keller MC. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. Am J Psychiatry 2011;168:1041–1049.
27. Angrist, J, Pischke J. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J Econ Perspect. 2010;24:3–30.
28. Imbens G, Rubin D, Sacerdote B. Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players. Am Econ Rev. 2001;91:778–794.
29. Sacerdote B. How large are the effects from changes in family environment? A study of Korean American adoptees. Q J Econ. 2007;122:119–157.

Cite this article as: Bierut L, Cesarini D (2015) How genetic and other biological factors interact with smoking decisions. *Big Data* 3:3, 198–202, DOI: 10.1089/big.2015.0013.

Abbreviations Used

GWAS = genome-wide association study
SNP = single nucleotide polymorphism
KHP = Kavli Human Project
DNA = deoxyribonucleic acid
CPD = cigarettes per day
TAG = Tobacco and Genetics Consortium
GxE = gene x environment