Using structural differences between environments to better understand rGE

Benjamin W. Domingue
Jason Boardman

Institute of Behavioral Science
University of Colorado Boulder

1. Introduction

If a genetic polymorphism has uniformly negative ramifications for fitness, evolutionary theory suggests it should be bred out of a population. Given that many seemingly deleterious variants are relatively common, it has long been thought that the effects of particular genes may vary as a function of environment. For this reason (as well as others), the occurrence of certain genes may vary as a function of environment. Gene-environment correlations (rGE) exist when allele frequencies vary across discrete physical, social, or behavioral environments. These associations are particularly important for gene-environment interaction (GxE) research because, as others have argued, failure to address rGE can lead to incorrect conclusions about the relevance of GxE (Jaffee & Price 2007; Keller 2014).[1] Population geneticists have shown that there are allele frequency differences across different socially defined racial and ethnic groups, a phenomenon known as population stratification, an extreme version of rGE. Population stratification has been a concern in genome-wide association studies (GWAS) since their inception. A variety of techniques, from genomic control (Devlin et al., 2001) to the usage of principle components (Price et al., 2006) have arisen to help researchers deal with this potential confounder.

In the same way that population stratification threatened to undermine the results of GWAS, more subtle forms of gene-environment correlations may confound inferences of gene-environment interaction studies. Simply understanding how to interpret certain genetic phenomena may require us to understand how genes are distributed across environments. For example, Fowler et al. (2011) suggest that genes, in particular DRD2 and CYP2A6, may play some role in friendship preference. However, Boardman et al. (2012B) suggest that the mechanism is perhaps social stratification by genotype (e.g., evocative rGE) rather than preference based on genotype. That is, both groups concluded that friends are more likely than chance to have similar genotypes at these loci, but they differ in terms of their explanations. Whereas Fowler et al. (2011) emphasize the selection of friendships by individuals, the Boardman et al. (2012B) group believes that forces outside of an individual's control (e.g., their school tracking) are the primary mechanisms for sorting along genetic lines.

An understanding of gene-environment interactions and correlations, a crucial topic for demography as many population-based studies are now collecting genetic data, is premised upon an understanding of how genes are distributed across these environments and what social mechanisms (e.g., institutions) are implicated in this sorting process. Perhaps even more fundamentally, understanding gene-environment correlations tells us about how genetic differences structure, and are structured by, environmental and social differences between individuals. The majority of previous work on this topic has been in either the

[1] It should be noted that rGE refers to whether an *environment* is associated with genotype while GxE is a question of whether an *outcome* is influenced in an interactive way by both genes and environment. Thus, one could inquire about whether the "E" part of GxE may demonstrate rGE.

candidate gene or twin literature. Given the increased availability of genome-wide data, there is a need to reconsider rGE in the context of this new data type. As we shall demonstrate, understanding rGE in genome-wide data is a complicated task. We first document whether rGE exists before focusing on how structural differences between environments might be relevant. In particular, we focus on the concept of genetic exogeneity, by which we mean environments that are unassociated with genotype (how exactly we would measure such an association is a complicated question we tackle in subsequent sections). This study addresses several research questions. *First*, what evidence do we see for rGE across a number of environments? *Second*, do the environments demonstrate patterns of genetic exogeneity that are consistent with hypotheses regarding rGE for these environments? *Third*, what is the consequence of ignoring rGE in the G+E paradigm? *Finally*, we attempt to highlight certain discrepancies between the models that might reasonably be considered as generating mechanisms for translating genotype to phenotype versus the kinds of models that tend to be estimated in practice.

2. Background

2a. Types of rGE

Theoretically, genetic exogeneity (the complete absence of rGE) occurs when an environment is unassociated with allele frequency at a single locus or multiple loci. In a laboratory setting, one could imagine placing, at random, certain mice in a cage with plenty of food and water and other mice in a cage with scarcer nutritional resources. Some response in behavior could then be observed and there may be interest in whether responses are a joint product of genotype and environment (the level of available nutrition in their cage). The random assignment of mouse to environment makes it clear that the environment is genetically exogenous. Human environments, however, rarely occur at random. In observational settings, selection is well known to be a major problem in understanding subsequent responses. If genetically endogenous environments are of interest in GxE studies, then there is a possibility that this form of endogeneity will confound inference.

We now discuss three forms of environments: exogenous environments, confounded environments, and endogenous environments. Exogenous environments are truly exogenous of genotype in the sense of the mice example above. In all likelihood, they are quite rare outside of the laboratory (or without a natural experiment) because any environment that has any component of selection is likely to have some genetic influence given the range of human behaviors and traits that are heritable. Consider birth year. Initially, one may see no reason why there should be genetic associations with birth year. Yet, decisions with respect to fertility are related to the larger economic climate (e.g., Sobotka et al., 2011) and certain classes of individuals may be less sensitive to these economic swings. This could induce a genetic patterning across birth year. Birth year would then be either a confounded or genetically endogenous environment, the distinction having to do with the nature of the genetic patterning.

The distinction between a confounded environment and a genetically endogenous environment is subtle and possibly difficult to determine in practice. Confounded environments are those in which there is a measurable difference of allele frequencies for populations in two different environments but the association is unrelated to any relevant biology. For example, there are large numbers of ancestors of Italian immigrants on the urbanized Eastern seaboard of the US and large numbers of ancestors of Norwegian immigrants in the more rural Midwest. These migration patterns will be evident in an examination of principal components or the MAF or specific loci, but genotype is not the reason for the selection into the environment (rather, history is). In the context of GxE studies, earlier work described

similar phenomena as causal and non-causal GxE models (Boardman et al. 2012C). Here, the existence of rGE may denote some causal association but there is also the possibility that the correlation between genotype and environment, albeit real in consequences for statistical inference, is not due to an underlying biological mechanism. Rather, the correlation between genotype and environment is the result of larger external forces that shape human history.

An endogenous environment, on the other hand, is one that is due to biologically relevant differences, not those purely driven by population stratification, in genotype across environments. The three commonly hypothesized models of rGE (passive, evocative, and active) would all be examples of endogenous environments. The first model, passive rGE, occurs when genes are structured by forces external to an individual. For example, children inherit their genes from their biological parents but they also inherit their social environments. Passive rGE is the least pernicious threat to GxE research since it can be potentially controlled for with either family level data, which allows for a focus on between-sibling differences holding environment constant, or at the very least identified by a comparison of between-environment genetic differences. Evocative and active rGE are more problematic since they posit interactivity between the individual's genotype and environment. Evocative rGE suggests that environments respond to genotype (e.g., a relatively irritable child may evoke a relatively hostile or cold parenting environment) while active rGE suggests that genes influence selection of environments which may lead to subsequent phenotypes (e.g., individuals who are genetically more likely to smoke cigarettes may select into friendship groups in which people are more likely to smoke). At present, it is difficult to distinguish between these different flavors of rGE and this paper does not focus explicitly on doing so. This paper tries to solve the simpler problem of identifying when rGE is present. In the context of genome-wide data, even understanding this relatively simple issue is challenging.

2b. Models mapping genotype to phenotype

We now discuss the models commonly used in the research literature to map genotype to phenotype and their theoretical properties in the presence of rGE. Figure 1 contrasts two "G+E" models (emphasizing that the genetic and environmental contributions are orthogonal) of some outcome, O. M1 is the simplest possible scenario in that genetics, G, and environment, E, both contribute to the outcome, but the contributions are completely independent. In this scenario, E is genetically exogenous. In this paper, we consider outcomes, O, with the following characteristics:

- They are massively polygenic: there are a large number of causal variants,
- individual causal variants have small effects,
- dominant and epistatic effects are second-order phenomena that are safely ignored.

For such outcomes O, a reasonable data generating process (DGP) would be (where we omit offsets for simplicity)

$$O_i = \alpha E_i + \sum_j \beta_j \mathrm{SNP}_{ij} + \epsilon_i. \text{ (Eqn 1)}$$

Note the additivity of the SNP effects and the fact that the genetic and environmental effects are orthogonal. These assumptions are crucial from the perspective of genome-wide association studies (GWAS). Without these assumptions, the entire GWAS framework would need to be reconsidered. For example, the meta-analysic strategy of combining information from a variety of studies in which the respondents are in very different environments may be foolhardy if the effect of G depends upon E.

Assuming that M0 is the DGP, what types of models would be used to estimate the effects of genes on some outcome? One approach would be that of the polygenic score. For example, there is interest (e.g., Belsky et al., 2013; Domingue et al., 2014) in estimates of b from models of the form

$$O_i = b\sum_j \widehat{\beta_J}\mathrm{SNP}_{ij} + e_i. \text{ (Eqn 2)}$$

Estimates of b are associated with the amount of outcome variation explained by the SNPs implicated in a GWAS as being plausibly causal variants (the GWAS generate $\widehat{\beta_J}$ which are used to construct the polygenic score).[2] This will be referred to as the "score" approach. An alternative approach (in fact, the only possible approach without appropriate GWAS information) would be to estimate genetic influence on a trait via estimated genetic similarity as in GCTA (Yang et al., 2010, 2011). In that approach, a very different type of model is estimated:

$$\boldsymbol{O} = \boldsymbol{g} + \boldsymbol{\epsilon_O} \text{ (Eqn 3)}$$

where it is additionally assumed that

$$\boldsymbol{g} \sim \mathrm{MVNormal}[0, \sigma^2_{G,O}G]. \text{ (Eqn 4)}$$

Equations 3 and 4 (which we refer to as the GCTA approach) ignore the environment, but since the environment is genetically exogenous this should have no impact on the relevant estimates of variance components ($\widehat{\sigma^2_{G,O}}$ and $\widehat{\sigma^2_{\epsilon,O}}$). A heritability estimate is then produced as

$$\frac{\widehat{\sigma^2_{G,O}}}{\widehat{\sigma^2_{G,O}} + \widehat{\sigma^2_{\epsilon,O}}}$$

which is a comparison of the genetic-specific variance of the outcome to its overall variance. It's imperative to note that Eqn 3 is estimated due to a lack of information (specifically about $\widehat{\beta_J}$ from Eqn 2), not because it is believe to be the true DGP. This GCTA model draws on the rich tradition of "animal models" (e.g., Wilson, 2010), which are widely used when pedigrees (that is, exact relationships between individuals) are known in fields such as plant and animal breeding.

M1 makes the helpful assumption that the environment is genetically exogenous. But this assumption is probably rarely true in real-world settings. Let's now suppose that M2 is the true scenario. In terms of Figure 1, what functional form does the top path take? If E is genetically endogenous rather than merely confounded, then the true DGP might be something similar to the weighted sum of risk alleles shown in Eqn 1. But that is a big "if". Given that (a) it may be virtually impossible to obtain reliable information on which variants are associated with specific environments given the measurement difficulties inherent in GWAS and (b) such associations may fluctuate quite strongly over time and place (compared to the perhaps more stable associations such as those that exist between genes and a trait such as height), it might be that the GCTA approach is actually a more reasonable approximation of how genetic patterning across environments develops: genetic similarity induces environmental similarity through a

[2] It's worth noting that the sample sizes currently available for GWAS are sufficient to produce only relatively noisy estimates of $\widehat{\beta_J}$. Thus, polygenic scores do not yet predict the amount of variance in outcomes such as BMI that we might expect given their heritability as estimated by something like GCTA (which also ignores dominant and epistatic effects).

variety of different biological mechanisms (perhaps shifting across time, place, and person) without there necessarily being a specific set of causal variants that we could reasonably hope to identify. Thus, we suppose that

$$\boldsymbol{E} = \boldsymbol{g} + \boldsymbol{\epsilon_E} \text{ (Eqn 5)}$$

where $\boldsymbol{g}$ is distributed as

$$\boldsymbol{g} \sim \text{MVNormal}[0, \sigma^2_{G,E} G]. \text{ (Eqn 6)}$$

O is then defined as in Eqn 1.

If attempts were then made to estimate Eqn 2, there would be a potential for bias in $\hat{b}$, given the ignored association between G and O. However, the nature of the bias is difficult to decipher since the causal variants underlying the G→E link are unspecified (or even poorly defined). Alternatively, if one were to estimate a GCTA model that ignores environment, then the estimated heritability is likely to be an overestimate of the direct genetic effect due to the indirect effect from genotype, to environment, to outcome. This is due to the fact that GCTA is recovering both the direct influence of genes (G→O) and the indirect influence of genes (G→E→O). We demonstrate that this is indeed the case via simulation in Section 4.2.

3. Methods

3a. Data

Empirical results are based on data from 8,487 non-Hispanic white respondents in the Health and Retirement Study (HRS, data came from the RAND Fat Files) born between 1920 & 1953 (inclusive). We focus on non-Hispanic whites since the genetic similarity estimates which are the basis for GCTA heritability estimates are sensitive to allele frequency differences that exist between those from different ancestral groups. Descriptive statistics and information on how the variables were computed is included in Table 1. We divide the variables into 3 sets:

- Environments: birth.year, birth.month, father.edu, mother.edu, veteran, urban.
- Behaviors: own.edu, smokev, drinkn, log.income, loneliness, migrant, num.kids.
- Phenotypes: height, weight, cognitive, bmi, iadla, cesd, self.health, num.conditions.

One crucial thing to note is that the migrant variable is defined as a person who has lived in multiple census divisions and is thus an indicator of a domestic migrant rather than an international migrant. In addition, the veteran status indicator has a large number of missing values because we only consider male veterans (to capitalize on the fact that the Selective Service Act only applies to males). Table 2 contains correlations between environments (the columns) and phenotypes and behaviors (the rows).

We pause to unpack some theoretical notes about the nature of the environments in question with respect to whether they will exhibit rGE. For some of the environments, it is easy to wager about their rGE status. Parental education should be genetically endogenous given that parents and offspring share genes associated to educational attainment (Conley et al., 2015). We assume that birth month should be genetically exogenous since there is little reason to suspect differences across month (although this is an

empirical question).[3] Birth year, veteran status, and urbanicity are more complicated. Birth year and veteran status should be largely genetically exogenous although each could be influenced by genes. As previously mentioned, birth year may be tied to larger economic trends. Service in the military was due to random processes for large numbers of men in the birth cohorts we use, but this was not universally true and there still may be important differences (perhaps in terms of health) for those men who do and do not serve. Urbanicity could be argued to have elements of both a confounded environment–due to ethnicity differences in the US between urban and rural residents–and an endogenous environment–some individuals may choose to live in an urban or rural location due to personality traits. One assumption that seems reasonable to make about these three environments (urbanicity, birth year, and veteran status) is that urbanicity should show more signs of rGE than the other two.

The phenotypes that we consider are either health-related or anthropometric phenotypes measured during the data collection process of HRS (e.g., present-day BMI is being reported, not a retrospective BMI in the respondent's youth). The behaviors are not as straightforward as many of these variables contain information about the past. For example, an individual's formal education is presumably long finished by the time they are enrolled in the HRS. By the time that an individual is retired, their level of education may be considered an environment in some respects (e.g., it is predictive of the SES of their friends and associates at time of retirement). We have used this rationale in previous work, for example, while studying the genetics of BMI (Boardman et al., 2014a). That study included family-based controls in an attempt to reduce confounding due to rGE, but it is possible that these were inadequate. This point is made to emphasize that there is likely to be an especially soft division between behaviors and environments. More educated people typically associate with more educated people. Smokers associate with other smokers. The exact implications of this porous boundary aren't straightforward but could have serious implications for some GxE research.

Genetic data for the HRS is based on DNA samples focus on single nucleotide polymorphisms (SNPs) collected in two phases. The first phase was collected via buccal swabs in 2006 using the QuiagenAutopure method. The second phase used saliva samples collected in 2008 and extracted with Oragene. Genotype calls were then made based on a clustering of both data sets using the Illumina HumanOmni2.5-4v1 array. SNPs are removed if they are missing in more than 5% of cases, have low MAF (0.01), and are not in HWE ($p<0.001$). We retained 1,698,845 SNPs after removing those which did not pass the QC filters. We also computed principal components (PCs) within the sample of non-Hispanic whites since there is evidence to suggest population stratification even in groups with a common ancestry (Nelis et al., 2009).

3b. Analyses

We estimate heritabilities using GCTA (Yang et al., 2010, 2011) for the outcomes discussed in Section 3a. We also make use of the fact that GCTA allows for linear predictors to control for principal components. We also conduct a simulation (described in detail in Section 4b) to examine the behavior of GCTA estimates when relevant environmental variation in genetics is ignored.

4. Results

[3] Although we suspect that there is little in the way of "information" in terms of birth month, it may still have implications for development (e.g., Stebelsky, 1991).

4a. Detecting rGE

Figure 2 contains heritability results (raw in black, adjusted for top 20 PCs in red) for phenotypes, behaviors, and environments. The raw estimate of height approaches the 0.4 estimate from the original GCTA paper (Yang et al., 2010). Of the behaviors, own education is estimated to have a heritability of 0.25 after controlling for PCs.[4] Smoking and log income are next with heritabilities of 0.21 and 0.2. Of the phenotypes, IADLA and cognitive ability have the lowest heritability. With only one exception, standard errors for all estimates in black are between 0.05 and 0.06 (the SE for veteran status is 0.12; it is larger due to the fact that we computed this based on only the males and thus have a decreased sample size). Standard errors for estimates in red are between 0.04 and 0.06. Since the standard errors are so consistently similar, we do not include them in Figure 2 but interpretation (with the exception of veteran status) is relatively straightforward: Any estimate around 0.1 or below is not statistically significant.

Several facts about controlling for PCs are worth note. For the behaviors, controls for the PCs uniformly lead to declines in the estimated heritabilities. The story is more complex for behaviors and environments. Heritabilities generally decline after controlling for PCs, but that is not the case for smoking (going from 0.21 to 0.24) and migration (0.06 to 0.09). For behaviors, we see large increases for birth year and veteran status. An increase in heritability after controlling for PCs would suggest that population stratification is leading to a type of Simpson's paradox. Although this might be theoretically possible, we tend to view such increases with skepticism.

Raw heritabilities for the environments (in black) are in two clusters. Urban environment and parental education have relatively large heritabilities (>0.25). Given that parents and offspring share a genetic predisposition towards educational attainment (Conley et al., 2015), the parental education results are not surprising. To further test this finding, we residualized parental education based on offspring education and estimated the GCTA heritability for residualized maternal and paternal education. We estimated heritabilities of 0.28 and 0.16 (SE=0.05 for both) for residualized maternal and education respectively. Thus, parental education net of one's own still seems to be heritable. While we anticipated some amount of genetic association with urban residence, the fact that heritability of urban environment is so large is somewhat surprising. One theory would be that this is due to population stratification within the non-Hispanic white sample. However, the result controlling for the top 20 PCs is nearly as large as the raw estimate, so perhaps this is not the case. Birth year, veteran status, and birth month have heritabilites below 0.1. The result for birth month is unsurprising. While the baseline results for birth year and veteran status make some amount of theoretical sense, it is interesting to note that these estimates increase when we control for PCs.

4b. Implications of ignoring rGE

We consider the implications of ignoring rGE under the G+E paradigm discussed above in two ways. First, we consider the implications of ignoring a genetically endogenous environment when estimating

[4] In previous work (Boardman et al., 2014), we noted a larger estimate for the heritability of education (0.33 after controlling for PCs). This was based on a different set of markers and respondents from HRS than what is used here. Both estimates are in the range of results from other sources. Rietveld et al. (2013) show estimates of 0.36 (the QIMR cohort) and 0.23 (the STR cohort). Marioni (et al., 2014) shows an estimate of 0.21 using the Generation Scotland cohort.

GCTA heritability. Second, we examine the empirical implications for outcomes where (1) an environment is associated with the outcome and (2) the environment has estimated non-zero heritability. We begin by discussing results of a simulation. We first simulated an environment via the GCTA model using the estimated genome-wide similarities computed for all respondents and choice of $\sigma^2_{G,E}$ (the error variance, $\sigma^2_{\epsilon,E}$, is fixed at 1). We then construct the purely genetic component of the outcome via the risk score model based on drawing causal effects for 50,000 SNPs from the standard normal distribution.[5] Both the environment and the purely genetic component are then standardized to have SDs of 1. Finally, the outcome is computed as the sum of (1) the purely genetic component, (2) the environment, and (3) error (with variance σ^2_e). Both $\sigma^2_{G,E}$ and σ^2_e are random draws from the uniform distribution on [0,2].

Figure 3 shows the results for 25 iterations of the simulation. The x-axis shows the true ratio (where the relevant quantities are known, not estimated) of $\mathrm{Var}(g_i)/\mathrm{Var}(O_i)$, where g_i are the individual entries of $\boldsymbol{g}$ from Eqn 5 (estimates of these quantities are of interest in plant and animal breeding, see Kruuk, 2004 for thoughts on estimation of these parameters in the presence of substantial environmental variation). This is the direct genetic influence of genotype on phenotype. The y-axis shows the estimated GCTA heritability when we ignore the environment. The dashed black 45 degree line shows where we would expect GCTA heritability estimates in the absence of the endogenous environment. However, when we ignore the endogenous environment, we consistently overestimate the true genetic contribution to phenotypic variance (by an average of 37%).

We also conduct an empirical test of whether ignored environmental associations tended to upwardly bias heritability estimates of behaviors by examining the estimated heritability of own education, drinking, and log income after considering the influence of urban environment. The top row of Table 3 contains heritability estimates for own education, drinking, and log income (the raw estimates in black from Figure 2). The second row contains estimates based on including the urban indicator as a predictor via GCTA. The third row contains estimates based on residualizing the outcome using the urban indicator. Rows 2 and 3 are generally quite similar and show slight declines in the estimated heritability as compared to the estimates from row 1. Although the reductions may seem slight, they are meaningful compared to the lack of change shown in the last row (italicized) in which GCTA models were estimated using the uninformative birth month variable as a control. We discuss the implications of this research in the subsequent section.

5. Discussion

Figure 2 suggests that some environments–specifically urbanicity and parental education–are genetically endogenous. This was as hypotheiszed, but it is interesting to note that parental education even net of offspring education continues to be heritable. This suggests that we are not picking up simply the fact that parents and offspring share half of their genetic material. The raw results for birth year and veteran status suggest that they are largely genetically exogenous. The results based on controlling for PCs are more complicated. Given the theoretical reasons to suspect that these variables are largely genetically exogenous, we are skeptical about the findings based on the PCs at present. The most important

[5] The normal distribution is possibly an inappropriate model for causal effects given its thin tails, but GCTA seems relatively robust to different types of genetic architecture (Speed et al., 2012).

findings are that we do see some support for our hypotheses about genetic exogeneity and endogeneity being present in the data.

What, exactly, are the implications of Figure 3? Why would the consistent overestimation of heritability be problematic? We believe that most environments are genetically endogenous and most phenotypes and behaviors are jointly influenced by both genes and environments. Thus, GCTA estimates of heritability may frequently be capturing the indirect pathway (G→E→O)through which genes influence environments and outcomes. The G→E pathway may be especially problematic given that it may be due to population stratification. Thus, GCTA estimates which are interpreted as explaining the variance associated with the direct G→O pathway may substantially overestimate the role that specific causal variants have directly on an outcome. Table 3 demonstrates this directly via a reduction in the estimated heritability in own education, drinking, and logged income when adjusted for urbanicity. Although the declines aren't huge (and are consistent with similar earlier work, Conley et al., 2014), we emphasize that they are due to controlling for only a single environment. Considering the range of possible environments which may be both genetically endogenous and causally implicated in these three traits, it is possible that direct GCTA estimate of these traits may substantially overestimate their actual heritability.

If one believes Turkheimer's (2000) "first law of behavior genetics" (all human behavioral traits are heritable), then there is a reason to ask why we are concerned about heritability estimates in the first place? There are many potential answers to that question and it would be encouraging to see more research attempt to understand the role of environment in their analyses. As one illustration, we critique an earlier study of our own (Boardman et al., 2014b) in which we asked whether education and various measures of health perhaps share a common genetic origin. While we attempted to minimize the role of population stratification through the use of principal components in our GCTA analyses, the results here suggest perhaps that such an approach may be difficult to interpret. One could imagine asking whether the results are robust to more stringent approaches. Programs like admixture (Alexander et al., 2009) or structure (Pritchard et al., 2000) could be used to extract a more ethnically homogenous group. Environmental factors are virtually impossible to rule out via GCTA, but we could also reconsider whether polygenic scores (if available) show the same pattern of correlations. In particular, would we see a correlation between a score for educational attainment and a score for depression? Would we fail to find a correlation between a score for educational attainment and BMI? Such inquiries would help us to ensure that we were accurately identifying pleiotropic effects (the goal of that research) rather than merely picking up on subtle environmental confounding that is common to both traits.

5b. Implications for GxE Research

This study has focused on the G+E paradigm. For social scientists, the GxE paradigm is much more interesting since it explicitly offers a mechanism through which environments might moderate the genetic influence on a trait. The candidate gene literature focused on interactions between variants of the candidate gene and environments. A classic example in this literature is the work of Caspi et al. (2003) in which a variant of 5-HTTT is shown to moderate the influence of life stress on depression. The twin literature, on the other hand, has seen focus on environments moderating the overall heritability of

a trait, Turkheimer et al's (2003) work on the heritability of IQ as a function of SES being a classic example. These are different ways of operationalizing GxE. While some work has attempted to blend these views (Boardman et al.'s (2012A) work on the influence of APOE as a function of neighborhood context being one example), it is worth carefully considering the implications of these different models in the context of genome-wide data.

Suppose that an outcome has a genetic component generated by the risk score model. The most intuitive formulation would be a GRSxE data generating model:

$$O_i = \alpha E_i + \sum_j \beta_j \mathrm{SNP}_{ij} + \gamma E_i \sum_j \beta_j \mathrm{SNP}_{ij} + \epsilon_i. \text{ (Eqn 7)}$$

There is already research attempting to estimate γ (e.g., Qi et al., 2012; Li et al., 2010; Meyers et al., 2013).[6] While this might be worthwhile research, there is a tough question easily obscured about the relationship between the causal variants (SNP_{ij}) and the environment. If there is a correlation between $\sum_j \beta_j \mathrm{SNP}_{ij}$ and E_i, then a significant estimate of the interaction, $\hat{\gamma}$, can result even if the DGP for the outcome O is Eqn 7 with $\gamma = 0$. Thus, trying to understand rGE should be a necessary starting point for any GxE study.

[6] It's worth noting that Qi et al. (2012) and Li et al. (2010) are examining what this paper would frame as behaviors: intake of sugary beverages and physical activity.

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome research, 19(9), 1655-1664.

Belsky, D. W., Moffitt, T. E., Sugden, K., Williams, B., Houts, R., McCarthy, J., & Caspi, A. (2013). Development and evaluation of a genetic risk score for obesity. Biodemography and social biology, 59(1), 85-100.

Boardman, J. D., Barnes, L. L., Wilson, R. S., Evans, D. A., & de Leon, C. F. M. (2012A). Social disorder, APOE-E4 genotype, and change in cognitive function among older adults living in Chicago. Social Science & Medicine, 74(10), 1584-1590.

Boardman, J. D., Domingue, B. W., & Fletcher, J. M. (2012B). How social and genetic factors predict friendship networks. Proceedings of the National Academy of Sciences, 109(43), 17377-17381.

Boardman, J. D., Roettger, M. E., Domingue, B. W., McQueen, M. B., Haberstick, B. C., & Harris, K. M. (2012C). Gene–environment interactions related to body mass: School policies and social context as environmental moderators. Journal of theoretical politics, 24(3), 370-388.

Boardman, J. D., Domingue, B. W., Blalock, C. L., Haberstick, B. C., Harris, K. M., & McQueen, M. B. (2014a). Is the gene-environment interaction paradigm relevant to genome-wide studies? The case of education and body mass index. Demography, 51(1), 119-139.

Boardman, J. D., Domingue, B. W., & Daw, J. (2014b). What can genes tell us about the relationship between education and health?. Social Science & Medicine.

Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., ... & Poulton, R. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science, 301(5631), 386-389.

Conley, D., Siegal, M. L., Domingue, B. W., Harris, K. M., McQueen, M. B., & Boardman, J. D. (2014). Testing the key assumption of heritability estimates based on genome-wide genetic relatedness. Journal of human genetics, 59(6), 342-345.

Conley D, Domingue BW, Cesarini D, Dawes C, Rietveld CA, Boardman JD. (2015). Is the Effect of Parental Education on Offspring Biased or Moderated by Genotype? Sociological Science. 2: 82-105.

Devlin, B., Roeder, K., & Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. Theoretical population biology, 60(3), 155-166.

Domingue, B. W., Belsky, D. W., Harris, K. M., Smolen, A., McQueen, M. B., & Boardman, J. D. (2014). Polygenic risk predicts obesity in both white and black young adults. PloS one, 9(7), e101596.

Fowler, J. H., Settle, J. E., & Christakis, N. A. (2011). Correlated genotypes in friendship networks. Proceedings of the National Academy of Sciences, 108(5), 1993-1997.

Jaffee, S. R., & Price, T. S. (2007). Gene–environment correlations: a review of the evidence and implications for prevention of mental illness. Molecular psychiatry, 12(5), 432-442.

Keller, M. C. (2014). Gene× environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. Biological psychiatry, 75(1), 18-24.

Kruuk, L. E. (2004). Estimating genetic parameters in natural populations using the 'animal model'. Philosophical Transactions of the Royal Society B: Biological Sciences, 359(1446), 873-890.

Li, S., Zhao, J. H., Luan, J. A., Ekelund, U., Luben, R. N., Khaw, K. T., ... & Loos, R. J. (2010). Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study. PLoS medicine, 7(8), e1000332.

Marioni, R. E., Davies, G., Hayward, C., Liewald, D., Kerr, S. M., Campbell, A., ... & Deary, I. J. (2014). Molecular genetic contributions to socioeconomic status and intelligence. Intelligence, 44, 26-32.

Meyers, J. L., Cerdá, M., Galea, S., Keyes, K. M., Aiello, A. E., Uddin, M., ... & Koenen, K. C. (2013). Interaction between polygenic risk for cigarette use and environmental exposures in the Detroit neighborhood health study. Translational psychiatry, 3(8), e290.

Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., ... & Metspalu, A. (2009). Genetic structure of Europeans: a view from the North–East. PloS one, 4(5), e5472.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics, 38(8), 904-909.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.

Qi, Q., Chu, A. Y., Kang, J. H., Jensen, M. K., Curhan, G. C., Pasquale, L. R., ... & Qi, L. (2012). Sugar-sweetened beverages and genetic risk of obesity. New England Journal of Medicine, 367(15), 1387-1396.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., ... & McMahon, G. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science, 340(6139), 1467-1471.

Sobotka, T., Skirbekk, V., & Philipov, D. (2011). Economic recession and fertility in the developed world. Population and development review, 37(2), 267-306.

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. The American Journal of Human Genetics, 91(6), 1011-1021.

Stebelsky, R. H. B. G. (1991). "Born to play ball" The relative age effect and Major League Baseball. Sociology of Sport Journal, 8, 146-151.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. Current Directions in Psychological Science, 9(5), 160-164.

Turkheimer, E., Haley, A., Waldron, M., d'Onofrio, B., & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. Psychological science, 14(6), 623-628.

Wilson, A. J., Reale, D., Clements, M. N., Morrissey, M. M., Postma, E., Walling, C. A., ... & Nussey, D. H. (2010). An ecologist's guide to the animal model. Journal of Animal Ecology, 79(1), 13-26.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., ... & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature genetics, 42(7), 565-569.

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics, 88(1), 76-82.

Table 1. Descriptive statistics for HRS variables used in study.

Var Name	Min	Max	Mean	SD	# NA	Notes
own.edu	0.00	17.00	13.19	2.53	37	HRS variable raedyrs
birth.year	1920.00	1953.00	1937.62	8.83	0	HRS variable rabyear
birth.month	1.00	12.00	6.52	3.44	0	HRS variable rabmonth
father.edu	0.00	17.00	9.93	3.53	889	HRS variable rafeduc
mother.edu	0.00	17.00	10.34	3.03	587	HRS variable rameduc
veteran	0.00	1.00	0.59	0.49	4809	HRS variables ravetrn, for males only
height	1.37	2.26	1.70	0.10	0	Max height over all waves
weight	38.94	174.29	79.21	17.37	3	Mean weight over all waves
cognitive	7.00	35.00	26.41	3.60	99	Max cognitive functioning over all waves (based on rXcogtot variables)
bmi	15.67	61.23	27.48	5.03	3	Mean BMI over all waves
iadla	0.00	3.00	0.26	0.62	0	Max IADLA over all waves
cesd	0.00	8.00	1.20	1.33	0	Mean CESD over all waves
self.health	1.00	5.00	2.55	0.86	0	Mean self-reported health over all waves
num.conditions	0.00	8.00	2.47	1.50	0	Max of rXconde variables which tabulate number of diagnosed conditions (high BP, diabetes, cancer, lung disease, heart problems, stroke, mental health problems, arthritis)
smokev	0.00	1.00	0.57	0.49	29	Whether a person ever reports having been a smoker
drinkn	0.00	7.57	0.71	0.98	1	Mean number drinks when drinking over all waves
log.income	6.40	15.72	10.84	0.77	0	Log of mean total income (rXitot variables)
loneliness	0.00	1.00	0.12	0.21	0	Mean of variables indicating whether a respondent felt lonely (rXflone)
urban	0.00	1.00	0.42	0.49	103	Respondent lived in an urban environment in 2008 (Based on HRS variable urbrur08).
migrant	0.00	1.00	0.66	0.47	5	Whether a person reports living in more than one census division
num.kids	0.00	14.00	2.61	1.64	12	HRS variable raevbrn

Table 2. Correlations between environments (columns) and phenotypes and behaviors (rows).

	birth.year	birth.month	father.edu	mother.edu	veteran	urban
height	0.05	-0.01	0.07	0.10	-0.06	0.00
weight	0.17	0.00	0.01	0.05	-0.09	-0.04
cognitive	-0.08	0.01	0.14	0.12	0.05	0.08
bmi	0.16	0.01	-0.04	-0.01	-0.07	-0.06
iadla	-0.18	0.00	-0.11	-0.09	0.03	-0.02
cesd	0.04	-0.01	-0.13	-0.11	-0.04	-0.02
self.health	-0.10	-0.02	-0.20	-0.19	0.00	-0.08
num.conditions	-0.26	-0.02	-0.18	-0.16	0.13	-0.05
own.edu	**0.16**	**0.02**	**0.40**	**0.39**	**0.00**	**0.13**
smokev	0.00	0.01	0.00	-0.01	0.14	0.04
drinkn	**0.15**	**0.02**	**0.14**	**0.15**	**0.00**	**0.11**
log.income	**0.33**	**0.02**	**0.30**	**0.31**	**-0.12**	**0.13**
loneliness	-0.05	-0.01	-0.10	-0.10	-0.03	-0.01
migrant	0.01	0.00	-0.10	-0.06	-0.02	-0.16
num.kids	-0.22	0.01	-0.10	-0.11	0.06	-0.07

Table 3. Heritability results for key behaviors net of environmental controls.

	Own education	SE	Drinkn	SE	log Income	SE
GCTA Raw	0.295	0.053	0.181	0.051	0.208	0.054
GCTA, control for urban	0.268	0.054	0.156	0.052	0.18	0.054
GCTA on residualized outcome	0.264	0.054	0.153	0.052	0.178	0.054
GCTA, control for birth month	*0.295*	*0.053*	*0.182*	*0.051*	*0.208*	*0.054*

Figure 1. G+E models for mapping genotype to phenotype.

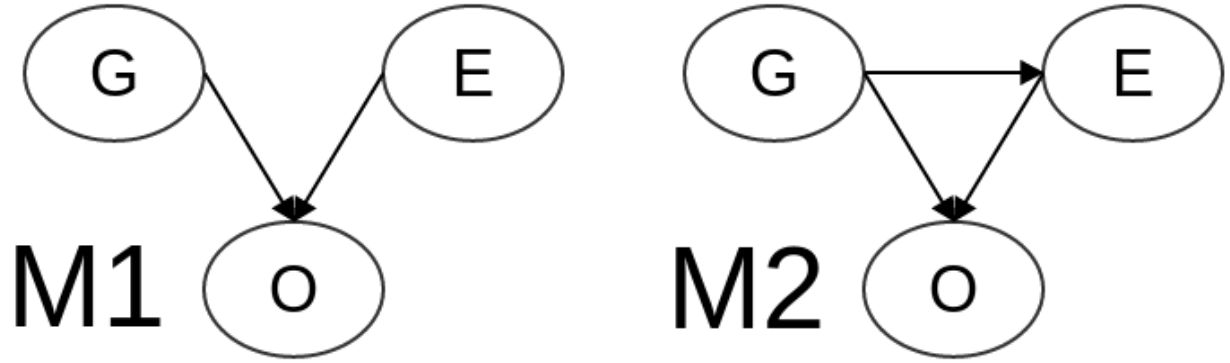

Figure 2. Heritability estimates, both raw and with controls for top 20 PCs, for phenotypes, behaviors, and environments.

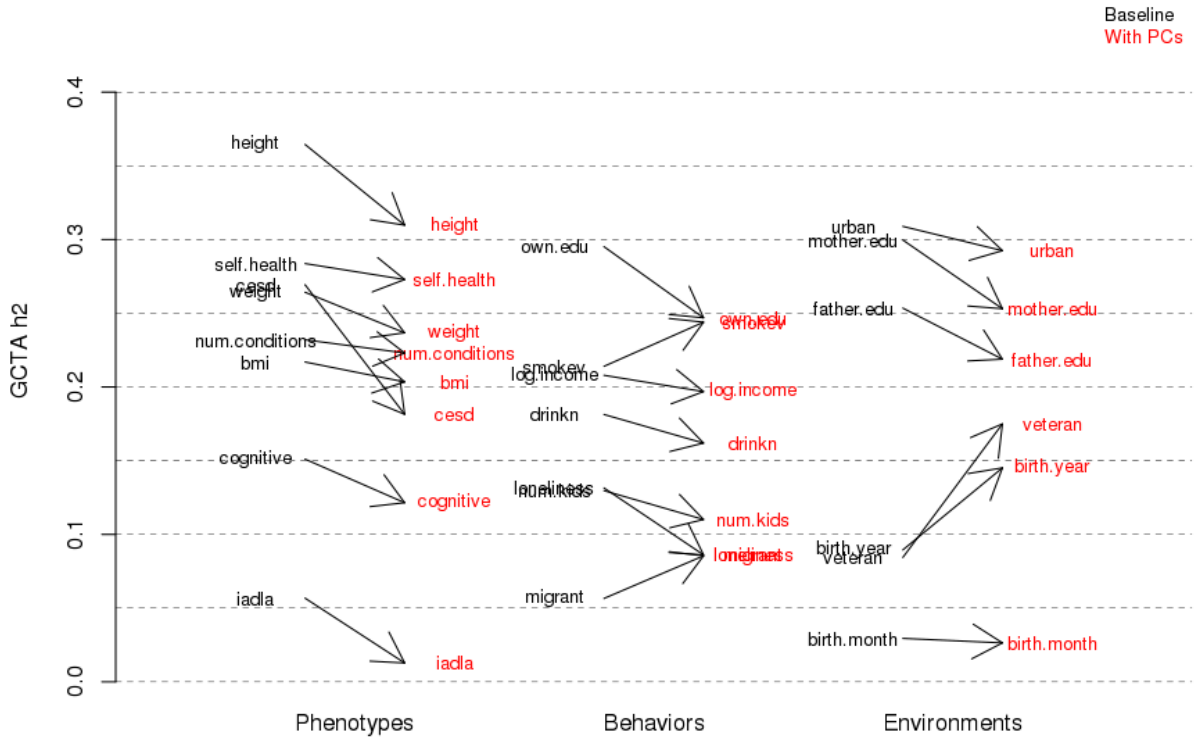

Figure 3. Comparison of “True h2” (the direct G→O link from M2) to GCTA h2 when M2 is the true DGP.