Understanding the role of social and economic factors in GCTA heritability estimates

David H Rehkopf, Stanford University

School of Medicine, Division of General Medical Disciplines

1265 Welch Road, MSOB 3rd Floor, MC 5411

Stanford, CA 94305 USA

650.725.0356

drehkopf@stanford.edu

Extended Abstract prepared for the Integrating Genetics in the Social Sciences Meeting 2014

Draft date: June 1, 2014

Abstract

A primary question in the social and biological sciences is parsing the proportion of phenotypic traits determined by the environment with that determined by genetic composition. In recent years, an approach using genome wide DNA sequence relationships between individuals has allowed this question to be approached in population level data, an approach known as genome-wide complex trait analysis (GCTA). The framing of results using this approach are that the estimates represent the proportion of a phenotypic trait that can be explained by genetic factors. This interpretation, however, ignores the extent to which genotypes of individuals may be correlated with observed social traits and geographic location. In order to understand the extent to which these factors may contribute, I performed SNP heritability using GCTA analysis conditional on potentially relevant social and economic factors. I first examined this question using the phenotype of height, replicating the heritability estimates in Yang and Visscher's original GCTA analysis in the Health and Retirement Study data, and then fit a model that was conditional on self-reported race/Ethnicity, level of education, level of income, parental level of education and geographic location. While there was essentially no attenuation of the heritability estimates of height, there was substantial attenuation for a measure of chronic disease and for work. I will subsequently extend this work to replicate all findings in the Framingham cohort.

A first order question in the social and biological sciences is parsing the proportion of phenotypic traits determined by the environment with that determined by genetic composition. At issue is, in the popular imagination at least, although perhaps quite erroneously, where the causal levers for intervention to improve health may exist. For most of the past century, the genetic versus environmental drivers of phenotypic traits have been examined through the use of twin studies, although estimates of these studies are subject to a number of potential biases for generalization beyond twin populations, they have given a window into heritability of complex traits (1). Nevertheless, one of the greatest limitations to these studies have been the limited sample sizes available, and questions as to whether findings could be extrapolated to the general population.

With the advent of the increase availability of data on large numbers of single nucleotide variations in large population databases a new approach has emerged to allow for an estimation of heritability (2). This approach has been termed genome-wide complex trait analysis (GCTA), and estimates of heritability are obtained through analysis of variance of all measured SNPs in relation to how they correlate between individuals with respect to a specific phenotypic trait. The seminal paper illustrating the utility of this approach estimated the heritability of height to be around 45%, while only 5% of the variance could be explained by SNPs identified in GWAS analyses (3).

However, the GCTA approach is limited by the fact that it does little to unpack where the heritability is coming from, a limitation noted earlier for previous approaches to measuring heritability by Jencks (4). The classic argument from Jencks is that to the extent that genotype impacts social or behavioral characteristics, these estimates of heritability therefore are more appropriately attributed to be the maximum possible to the extent that they correlate with social and economic factors that influence the phenotype. While from perhaps the most standard perspective there would be a violation of temporal causal ordering to Jencks's argument, that is, that genotype is prior thus the social and behavioral factors are on the causal pathway - there are at least two limitations to this criticism. The first is that from a purely practical and social policy perspective, the social and economic factors may be more amenable to interruption and change and thus be of greater potential interest. But secondly, that the presumed temporal ordering ignores how from a multi-generational perspective there could be potential social and economic forces causing shifts in genotype in the population over the course of multiple generations. While both of these counter-arguments remain for the most part speculative, they at least raise the issue of the limitations of heritability estimates that ignore what non-genetic influences are captured within those estimates.

In the following analysis I address this through conditioning on a small set of the most fundamental social and economic measures and compare these conditional to marginal heritability estimates as estimated through SNP relatedness implemented with GCTA.

The analysis uses data from the Health and Retirement Study (HRS), which began in 1992 with a nationally representative sample of non-institutionalized US residents born 1931-1941, and their spouses. I utilized phenotype and demographic data from the HRS Rand Files Version M that were cleaned and organized to facilitate longitudinal analysis across waves of data from 1992-2010. DNA was extracted from saliva samples that were collected from participants in 2006 and 2008. Linking individuals in the demographic dataset with genetic data resulted in an overall sample size of 98,932 observations on 12,845 individuals. The racial/ethnic distribution of the population was 10,527 non-Hispanic white, 1706 non-Hispanic blacks, 1176 Hispanics and 612 individuals that did not self-classify into one of these groups. At baseline there were 5,273 men and 7,572 women. The mean age at first observation was 57, and the overall mean age across all observations was 65. The minimum age was 24, and the maximum age was 107. There were 1524 individuals under the age of 50 because spouses of the main sample age 50 and above were included. At baseline, 37% of individuals were working full time, and 38% of individuals were in the labor force. Genotyping was coordinated by the Health and Retirement Study investigators using the Illumina 2.5-million SNP chip that covers all SNPs with a minor allele frequency of at least 5%. I accessed this data through dbGaP, and received permission from HRS to link the genetic identifiers to the detailed social and demographic data in the primary HRS survey. When constructing the genetic relatedness matrix I removed individuals where estimated relatedness is greater than 0.025. This corresponds to removing individuals related as closely as cousins 2-3 times removed. After removing those individuals the sample size was 8922.

I focused the current set of analyses of three contrasting phenotypes of interest. The first was height. Not only is height a well-characterized classic trait in terms of heritability, it also was the subject of the seminal GCTA paper (3), and thus I first replicate their results in HRS data. Secondly, I examine a measure of chronic disease which includes whether an individual ever had cardiovascular disease, cancer, diabetes or stroke. Future analyses will examine more refined health categories and biomarkers. Thirdly, I examined whether an individual was in the labor force at the baseline of the HRS survey.

I ran two sets of REML models with implementation in GCTA for each of these three outcomes. The first was a standard marginal analysis. The second analysis was fit conditional on region, race, parental education, own education and own earnings. For regional analysis I used the nine census regions available in the publically available data (Pacific, Mountain, West North

Central, East North Central, Middle Atlantic, New England, West South Central, East South Central and South Atlantic). Parental education of the subjects used the highest obtained level of education of either parent, or the only parent if on parent was missing, as a categorical variable of less than high school, high school or more than high school. These same categories were used for the subjects own level of education. Race/ethnicity was used as self-reported as white, black, Hispanic or other. Individual level earnings were taken from baseline enrollment in the HRS survey and used as quartiles.

The table below shows the results of the analyses of heritability based on genetic relatedness of individuals with shared phenotypes, with both the marginal and conditional estimates presented. For height, I estimate a level of heritability slightly higher than that previously estimated by Yang {Yang, 2010 #3250}, but of a generally similar magnitude. After refitting the model conditional on region, education, parents education, race/Ethnicity and earnings, results remained generally similar. Next, I ran the same set of models on the phenotype of chronic disease. While the heritability was modest in the marginal model, it was markedly attenuated in the conditional model, indicating that heritability estimates based on genetic relatedness are in large part explained by the geographic and social covariates. Finally, for labor force participation, there was only a small amount of heritability, but even this small amount was almost completely accounted for by variation with geographic and social covariates.

From the limited set of phenotypes I have up until now examined, the extent to which a small set of geographic and social covariates explain these is phenotype dependent. For a very specific physiologic trait, heritability estimates were not impacted, but for a very general health trait, and for a highly socially influenced trait, heritability estimates were greatly impacted. The findings have implications for the meaning of heritability estimates, in particular for analyses that are not investigating very specific physiological traits. The next steps are to examine a wider range of phenotypes, and examine the extent to which different selections of covariates impact attenuation of SNP based heritability.

Table

	standard GCTA models		with demographic controls	
source	variance	SE	variance	SE
HEIGHT				
V(G)	0.0069	0.0013	0.0057	0.0012
V(e)	0.0045	0.0009	0.0041	0.0008
Vp	0.0115	0.0004	0.0099	0.0004
V(G)/Vp	0.6028	0.0923	0.5807	0.1037
CHRONIC DISEASE				
V(G)	0.0283	0.0220	0.0143	0.0234
V(e)	0.1485	0.0157	0.1520	0.0167
Vp	0.1769	0.0077	0.1664	0.0080
V(G)/Vp	0.1603	0.1186	0.0862	0.1374
IN LABOR FORCE				
V(G)	0.0330	0.0312	0.0000	0.0145
V(e)	0.4686	0.0229	0.0966	0.0103
Vp	0.5017	0.0120	0.0966	0.0049
V(G)/Vp	0.0659	0.0609	0.0000	0.1499

Table Notes: demographic controls include own education, parental highest level of achieved education, self-reported race/Ethnicity, quartiles of earnings at baseline and census region (9 categories)

References

1. Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. Nature reviews genetics. 2002;3(11):872-82.
2. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76-82. doi: 10.1016/j.ajhg.2010.11.011. PubMed PMID: 21167468; PubMed Central PMCID: PMC3014363.
3. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9. doi: 10.1038/ng.608. PubMed PMID: 20562875; PubMed Central PMCID: PMC3232052.
4. Jencks C. Heredity, environment, and public policy reconsidered. Am Sociol Rev. 1980;45(5):723-36. Epub 1980/10/01. PubMed PMID: 7425434.