Hum Genet (2015) 134:259–267
DOI 10.1007/s00439-014-1519-0

ORIGINAL INVESTIGATION

Genetic and environmental components of family history in type 2 diabetes

Marilyn C. Cornelis · Noah Zaitlen · Frank B. Hu · Peter Kraft · Alkes L. Price

Received: 18 August 2014 / Accepted: 8 December 2014 / Published online: 30 December 2014

Abstract Family history of diabetes is a major risk factor for type 2 diabetes (T2D), but whether this association derives from shared genetic or environmental factors is unclear. To address this question, we developed a statistical framework that models four components of variance, including known and unknown genetic and environmental factors, using a liability threshold model. Focusing on parental history, we simulated case–control studies with two first-degree relatives for each individual, assuming 50 % genetic similarity and a range of values of environmental similarity. By comparing the association of parental history with T2D in our simulations to case–control studies of T2D nested in the Nurses' Health Study and Health Professionals Follow-up Study, we estimate that first-degree relatives have a correlation of 23 % (95 % CI 15–27 %) in their environmental contribution to T2D liability and that this shared environment is responsible for 32 % (95 % CI 24–36 %) of the association between parental history and T2D, with the remainder due to shared genetics. Estimates are robust to varying model parameter values and our framework can be extended to different definitions of family history. In conclusion, we find that the association between parental history and T2D derives from predominately genetic but also environmental effects.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-014-1519-0) contains supplementary material, which is available to authorized users.

M. C. Cornelis (✉)
Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, 680 N Lake Shore, Suite 1400, Chicago, IL 60611, USA
e-mail: mcorneli@hsph.harvard.edu; marilyn.cornelis@northwestern.edu

M. C. Cornelis · F. B. Hu
Department of Nutrition, Harvard School of Public Health, Boston MA, USA

N. Zaitlen
Department of Medicine, Lung Biology Center, University of California San Francisco, San Francisco, CA, USA

F. B. Hu · P. Kraft · A. L. Price
Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

P. Kraft · A. L. Price
Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

A. L. Price
Broad Institute of MIT and Harvard, Cambridge, MA, USA

Introduction

Diabetes is a leading cause of morbidity and mortality in developed countries. The worldwide prevalence of the disease is projected to double from the 371 million estimated in 2012 to 551 million in 2030, with type 2 diabetes (T2D) accounting for more than 90 % of these cases (International Diabetes Federation 2011). Roughly 40 % of the variance in T2D is due to genetics (Kaprio et al. 1992; Poulsen et al. 1999), suggesting that both genetic and environmental factors contribute substantially to an individual's T2D risk. Increased adiposity, inactivity, smoking, excessive caloric intake and poor diet quality have consistently been associated with risk of the disease (Hu 2011). Thus far, 5.7 % of the variance in T2D can be explained by known variants identified by genome-wide association studies (GWAS) (Morris et al. 2012). A positive family history of diabetes has been associated with a 2- to 30-fold increased risk of T2D (Abbasi et al. 2011; Hemminki et al. 2010; Lyssenko et al. 2008; Meigs et al. 2008; The InterAct Consortium 2013; van't Riet et al. 2010), but whether this association derives from genetic or shared environment is unclear.

Known environmental factors such as excess adiposity and, to a lesser extent, specific diet and lifestyle factors, reportedly explain up to 40 % of the association between having a family history of diabetes and risk of T2D (Abbasi et al. 2011; The InterAct Consortium 2013; van't Riet et al. 2010). Many of these factors are also partly genetic but how this contributes to the relationship between family history and the development of T2D is unclear (Rice et al. 1995). Known genetic variants for the disease explain very little of this association (Cornelis et al. 2009; Lyssenko et al. 2008; Meigs et al. 2008; The InterAct Consortium 2013), consistent with the weak correlation between family history of diabetes and these known variants (Vassy et al. 2011).

Early work of Khoury et al. (1988), which relied exclusively on simulations without analyzing real phenotypes, suggested that shared environment may not be responsible for the majority of shared familial risk of complex disease. Twin and family studies have been the standard approaches to investigating components underlying familial aggregation of a disease but are underpowered and subject to biases (Hopper et al. 2005; Kendler et al. 1993; Zaitlen et al. 2013; Zuk et al. 2012). The growing availability of population-based datasets of unrelated individuals with genetic, environmental and personal family history data offers a potentially powerful opportunity to understand family history of a disease. In the current study we quantify the extent to which family history of diabetes derives from genetic or shared environmental factors. We use a combination of simulated and empirically derived models, maximizing the availability of known genetic and environmental risk factors for the disease in unrelated individuals.

Materials and methods

Statistical model

Our statistical framework is based on the liability threshold model, involving an underlying normally distributed liability (Falconer 1967). An individual is a disease case if the liability exceeds a threshold T, where T is determined by the prevalence of disease in the population. We partition the liability into genetic (G) and environmental (E) components; both G and E are further partitioned into known (1) and unknown (2) components. Thus, the liability is the sum of G1 (known genetic effects), G2 (unknown genetic effects), E1 (known environmental effects) and E2 (unknown environmental effects), with liability-scale variances σ^2_{G1}, σ^2_{G2}, σ^2_{E1}, σ^2_{E2}, respectively. Because known "environmental" effects can be partly genetic (for example, the heritability of BMI is ~0.5 (Maes et al. 1997; Visscher et al. 2012), we further partition E1 into a genetic component E1G ($\sigma^2_{\mathrm{E1G}} = \sigma^2_{\mathrm{E1}}x$, where x is the proportion of E1 explained by genetics), and a genuinely environmental component E1E ($\sigma^2_{\mathrm{E1E}} = \sigma^2_{\mathrm{E1}} - \sigma^2_{\mathrm{E1G}}$). The variance explained by family history (FH) derives from G1, G2, E1G, E1E and/or E2. Our objective is to estimate how much of this is genetic (G1, G2 and/or E1G) and how much is environmental (E1E and/or E2). We assume that first-degree relatives have a correlation of 0.5 for each genetic component (G1, G2 and E1G) and a correlation of ξ for each environmental component (E1E and E2), where ξ is unknown.

Simulation framework

We first generated liabilities $\varphi \sim N(0, 1)$ for a given number of individuals (N) and then defined case–control phenotype π to equal 1 (case) if $\varphi \geq T$ or 0 (control) if $\varphi < T$. We used the threshold $T = 1.28$, which corresponds to a disease prevalence of 10 %. When modeling case–control ascertainment, we subsampled a user-defined number of cases and controls from the larger set of simulated individuals (see below).

Liabilities were simulated based on four components G1, G2, E1 and E2 with liability-scale variances σ^2_{G1}, σ^2_{G2}, σ^2_{E1}, σ^2_{E2}, respectively (see above). We used values of $\sigma^2_{\mathrm{G1}} = 0.03$ and $\sigma^2_{\mathrm{E1}} = 0.18$ obtained empirically (see "Results") based on variance explained by known genetic (σ^2_{G1}) and environmental (σ^2_{E1}) associations with T2D. We set $\sigma^2_{\mathrm{G2}} = 0.40 - \sigma^2_{\mathrm{G1}} = 0.37$ based on estimated heritability ($\sigma^2_{\mathrm{G1}} + \sigma^2_{\mathrm{G2}}$) of 0.40 for T2D (Kaprio et al. 1992; Poulsen et al. 1999), and thus $\sigma^2_{\mathrm{E2}} = 0.60 - \sigma^2_{\mathrm{E1}} = 0.42$. We also performed simulations in which known "environmental" effects (E1) were partitioned into a genetic component E1G ($\sigma^2_{\mathrm{E1G}} = \sigma^2_{\mathrm{E1}}x$, where $x = 0.5$, corresponding to the heritability of BMI (Maes et al. 1997), a major environmental risk factor for T2D) and a genuinely environmental component E1E ($\sigma^2_{\mathrm{E1E}} = \sigma^2_{\mathrm{E1}} - \sigma^2_{\mathrm{E1G}}$) (see above).

To simulate parental history (referred herein as 'family history' [FH]) of diabetes, we simulated exactly two first-degree relatives (i.e. mother and father) for each individual, assuming parent–offspring correlation of 0.5 for genetic components and ξ for environmental components, for various values of ξ (0.0–0.5, in increments of 0.01). Each individual was assigned FH = 1 if at least one of his/her two parents was a disease case, and FH = 0 otherwise.

We conducted simulations under both case–control ascertainment (proportion of disease cases oversampled) and random ascertainment (proportion of disease cases equal to prevalence 10 %). We note that the variance explained by FH may be part of G1, G2, El and/or E2, which together explain all of the phenotypic variance. For each value of ξ, we ran linear regressions to evaluate the proportion of phenotypic variance explained by any combination of G1, E1, and/or FH (the variables that are observed in a real study). The proportion of variance was estimated on the observed scale, but can be transformed to the liability scale (either with or without correction for case–control ascertainment, as appropriate) as described by Lee et al. (2011):

$$h_l^2 = h_{o_{cc}}^2 \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)},$$

[Eq. 23 in Lee et al. (2011)]

where h_l^2 is the proportion of variance estimated on the liability scale, $h_{o_{cc}}^2$ is the proportion of variance estimated on the observed scale under case–control ascertainment, K is the prevalence of disease, P is the proportion of cases in the study and $z = \frac{e^{\frac{-T^2}{2}}}{\sqrt{2\pi}}$. At a disease prevalence of 10 %, variance components on the observed scale with case–control ascertainment are 0.954 times those on the liability scale, and variance components on the observed scale without case–control ascertainment are 0.344 times those on the liability scale.

$\hat{\xi}$ can then be used in Eq. 1 to estimate the proportion of shared liability that is due to genetics ($\pi_{\text{FH,G}}$) vs. environment ($\pi_{\text{FH,E}} = 1 - \pi_{\text{FH,G}}$):

$$\hat{\pi}_{FH,G} = \frac{0.5(\sigma_{G1}^2 + \sigma_{G2}^2 + \sigma_{E1G}^2)}{0.5(\sigma_{G1}^2 + \sigma_{G2}^2 + \sigma_{E1G}^2) + \hat{\xi}(\sigma_{E1E}^2 + \sigma_{E2}^2)}, \quad (1)$$

which is the ratio between the genetic covariance between parent and offspring ($r_{\text{p-o,G}}$) and the total phenotypic covariance between parent and offspring ($r_{\text{p-o}}$); these concepts can be generalized to any relationship class.

Simulations described above were repeated under different assumed values of heritability (G1 + G2, 0.20–0.60 with increments of 0.05) and proportion of E1 explained by genetics (x, 0.0–0.75 with 0.05 increments). In a secondary analysis, we simulated a third first-degree relative (i.e. sibling) and repeated the entire set of experiments.

Nested case–control GWAS of type 2 diabetes

Empirical estimates of variances in T2D risk explained by FH and known genetic (G1) and environmental (E1) risk factors (and combinations thereof) were obtained from two case–control T2D GWAS nested in the Nurses' Health Study (NHS) and Health Professionals Follow-up Study (HPFS) (Qi et al. 2010) and were compared to simulations with the same number of cases and controls. The NHS was established in 1976 when 121,700 female registered nurses aged 30–55 years and residing in the US completed a mailed questionnaire on medical history and lifestyle characteristics (Colditz and Hankinson 2005). The HPFS was initiated in 1986 when 51,529 male health professionals between 40 and 75 years of age and residing in the US completed a questionnaire on lifestyle and medical history (Chu et al. 2001). Every 2 years, follow-up questionnaires have been sent to participants from both cohorts to update information on exposures and newly diagnosed disease. Details regarding the study design including population, data collection, assessment of T2D (see Supplementary text), T2D case–control selection, genotyping quality control and imputation have been reported elsewhere (Cornelis et al. 2009; Hu et al. 2001; Laurie et al. 2010; Manson et al. 1991; Qi et al. 2010). We restricted the current analysis to incident cases of T2D and to participants of European ancestry with complete genotyping data for T2D risk variants: 1,746 controls and 1,237 cases for NHS, 1,287 controls and 867 cases for HPFS. We replaced missing non-genetic covariate data (<9 % of NHS and <3 % of HPFS), with the study mean value of that covariate. All participants provided written informed consent, and the study was approved by the institutional review board of the Brigham and Women's Hospital and the Human Subjects Committee Review Board of Harvard School of Public Health.

Measures of family history of diabetes (FH)

In the NHS, women were asked to report whether any of their first-degree family members (father, mother, and/or siblings) ever had diabetes in the questionnaires mailed in 1982, 1988, and 1992. Corresponding data were collected for men (HPFS) in 1987, 1990, 1992 and 2008. We considered paternal, maternal or sibling history of diabetes to be positive if the respondent indicated so in any of the questionnaires and negative if responses were "no" or "missing". Parental history was defined based on a positive paternal and/or maternal history.

Measures of known environmental factors (E1)

For HPFS, anthropometrical and lifestyle factors was derived from the baseline (1986) questionnaires (Colditz et al. 1997; Rimm et al. 1991). For NHS, a more complete assessment of lifestyle factors was administered in 1980 and thus served as the baseline year for this cohort. BMI was calculated as weight (in kilograms) divided by the square of height (in meters). Physical activity was expressed as metabolic equivalent task (MET) hours of moderate to vigorous exercise per week (Wolf et al. 1994). METs were calculated for NHS in 1986 but correlated with the less precise measure of physical activity from 1,980 (h/week). A semi-quantitative food frequency questionnaire was used to derive an Alternative Healthy Eating Index (AHEI) score, a measure of overall diet quality that has previously been associated with T2D in these cohorts (Chiuve et al. 2012; Willett 1998).

Measures of known genetic factors (G1)

For each participant, a genetic risk score (GRS) for T2D was calculated by summing the values of 55 SNPs tagging reproducibly associated T2D loci reaching GW levels of significance in European populations (Morris et al. 2012). This SNP list did not include SNPs that were sex specific or that were identified in earlier GWAS but not confirmed in the recent meta-analysis (Rice et al. 1995; Rimm et al. 1991). rs9936386 (*FTO*), also associated with obesity, was included among the 55 SNPs. We assumed an additive genetic model for each SNP, applying a linear weighting of 0, 1, and 2 to genotypes containing 0, 1, or 2 alleles previously associated with increased risk, respectively (Cornelis et al. 2009). Each SNP was weighted according to its relative effect size (β coefficient) from the recent meta-analysis (combined discovery + replication stages) by Morris et al. (2012). The score was rescaled to reflect the number of risk alleles and theoretically ranges from 0 to 110, with higher scores indicating a higher genetic susceptibility to T2D.

Statistical analysis

The variances explained by each of the components G1, E1, and FH or combinations of these components in simulations (for each value of ξ, shared environment) were compared to their respective values in real data. The estimate of ξ for a sample/dataset ($\hat{\xi}$) is the value of ξ that minimizes the sum of squared residuals based on differences between observed variances in analyses of real phenotypes and the fitted variances in analyses of simulated phenotypes: $(\hat{\sigma}^2_{\text{G1}} - \sigma^2_{\text{G1}})^2 + (\hat{\sigma}^2_{\text{E1}} - \sigma^2_{\text{E1}})^2 + (\hat{\sigma}^2_{\text{FH}} - \sigma^2_{\text{FH}})^2 + (\hat{\sigma}^2_{\text{G1,E1}} - \sigma^2_{\text{G1,E1}})^2 + (\hat{\sigma}^2_{\text{G1,FH}} - \sigma^2_{\text{G1,FH}})^2 + (\hat{\sigma}^2_{\text{E1,FH}} - \sigma^2_{\text{E1,FH}})^2 + (\hat{\sigma}^2_{\text{G1,E1,FH}} - \sigma^2_{\text{G1,E1,FH}})^2$. We simulated 1,000 independent simulated data sets with sample sizes and parameter values matching those of the NHS and HPFS and computed estimates of ξ and corresponding non-parametric 95 % CI. Because BMI was the primary contributor to E1, and the narrow-sense heritability of BMI is widely believed to lie between 0.4 and 0.6 (Visscher et al. 2012), we computed separate 95 % confidence intervals for $x = 0.4$ and $x = 0.6$ and conservatively report the union of those confidence intervals as our primary result. For each genetic data set, case–control status and all independent variables were normalized [i.e. mean of 0 and standard deviation (SD) of 1] prior to modeling. E1 consisted of BMI (kg/m^2), current smoking (yes/no), physical activity (METs/week), alcohol consumption (g/day), diet quality (AHEI) and age. The latter, however, contributed little to T2D liability in this data set as a result of age-matching. Linear regression was used to obtain empirical estimates of the variance of T2D explained by any combination of G1, E1 and/or FH. Similarly, we obtained empirical estimates of the variance of T2D explained by any combination of E1 and/or FH using the entire NHS ($n = 84{,}880$) and HPFS ($n = 39{,}694$) cohort, for which G1 is not available. Our primary analysis included FH defined using parental history, but we also extend our framework to other definitions of FH.

Results

Baseline characteristics of cases and controls are presented in Table 1. In logistic regression analysis adjusting for age, lifestyle factors, BMI and the GRS, the odds ratio (95 % CI) of T2D associated with a positive parent history of diabetes was 2.82 (2.35–3.37) in the NHS and 2.86 (2.30–3.55) in the HPFS.

Table 2 presents the variances in T2D risk explained by FH and G1 and E1 risk factors (and combinations thereof) observed in the NHS and HPFS. A parental history of diabetes explained 0.07 of the variance in risk among NHS and 0.06 among HPFS. For both cohorts, the risk variance explained by G1 ($\hat{\sigma}^2_{\text{G1}}$) and E1 ($\hat{\sigma}^2_{\text{E1}}$) was approximately 0.03 and 0.18, respectively; thus $\sigma^2_{\text{G2}} = 0.37$ and $\sigma^2_{\text{E2}} = 0.42$. BMI was the primary contributor to $\hat{\sigma}^2_{\text{E1}}$: when considered alone, it explained 0.17 of the variance in T2D risk. Our estimate of G1 (3 %, based on 55 SNPs) in NHS and HPFS is lower than the 5.7 % (based on 63 SNPs) reported in the latest meta-analysis by Morris et al. (2012). The differences may be due to the number of SNPs analyzed, but are more likely due to differences between cohorts or to the winner's curse (Garner 2007). Supplementary Tables S1 (NHS) and S2 (HPFS) present equivalent data to that of Table 2 applying different definitions of FH. For both cohorts accounting for sibling history of the disease added little to the variance of T2D risk beyond that of parental history. All variances in T2D risk explained by FH and E1 when estimated from the ascertained case–control studies of these cohorts were roughly threefold greater than those when estimated from the full cohorts (Supplementary Tables S3 and S4), consistent with the fact that observed-scale variance increases under case–control ascertainment (0.954 vs. 0.344; see above) (Lee et al. 2011).

Results from simulations to estimate parental shared environment ($\hat{\xi}$) for these two studies (where $\sigma^2_{\text{G1}} = 0.03$, $\sigma^2_{\text{E1}} = 0.18$, $\sigma^2_{\text{G2}} = 0.37$ and $\sigma^2_{\text{E2}} = 0.42$) are displayed in Fig. 1. When the proportion of E1 explained by genetics (x) is 0.5, and thus $\sigma^2_{\text{E1G}} = 0.09$ and $\sigma^2_{\text{E1E}} = 0.09$, the value of ξ that best fits the observed data reported in Table 2 for NHS ($\hat{\xi}$) was 0.23 with 95 % CI (accounting for the uncertainty in x) of 0.16–0.27 (Fig. 1a). The proportion of shared liability due to environment ($\hat{\pi}_{\text{FH},E}$) is estimated to be 0.32 (95 % CI 0.25–0.37). Estimates for HPFS were consistent, but slightly lower (Fig. 1b). Averaging across cohorts, we

Table 1 Baseline characteristics of the Nurses' Health Study and Health Professionals Follow-up Study nested case–control studies of T2D

Characteristic	NHS			HPFS		
	Controls[a]	Cases[a]	P[b]	Controls[a]	Cases[a]	P[b]
	$n = 1{,}746$	$n = 1{,}237$		$n = 1{,}287$	$n = 867$	
Age	47.1 (6.8)	47.1 (6.8)	0.99	55.5 (8.4)	54.3 (8.3)	0.001
BMI, kg/m^2	24.2 (4.2)	28.3 (5.1)	<0.0001	25.0 (2.7)	28.0 (3.9)	<0.0001
Alternative Healthy Eating Index	31 (5.9)	31 (5.7)	0.01	48 (11)	45 (10)	<0.0001
Current Smoking, *n* (%)	340 (19)	322 (26)	<0.0001	136 (11)	138 (16)	0.0003
Alcohol, g/day	6.5 (9.6)	4.7 (8.9)	<0.0001	12.1 (15.1)	11.8 (16.3)	0.70
Physical activity, METs/week	14 (16)	12 (14)	<0.0001	21 (23)	14 (19)	<0.0001
Family history of diabetes, *n* (%)	418 (24)	630 (51)	<0.0001	362 (28)	460 (53)	<0.0001
Parent history[c]	358 (21)	558 (45)	<0.0001	267 (21)	381 (44)	<0.0001
Sibling history[c]	111 (6)	198 (16)	<0.0001	142 (11)	229 (26)	<0.0001
Number of siblings[d]	2.49 (1.99)	2.47 (1.97)	0.79	2.18 (1.83)	2.10 (1.78)	0.30
T2D genetic risk score (G1)	58.5 (5.0)	60.3 (5.1)	<0.0001	58.7 (5.1)	60.5 (4.9)	<0.0001

[a] Data are means (standard deviation) for continuous and *n* (%) for dichotomous variables

[b] Results from analysis of variance (continuous variables) or χ^2 (dichotomous variables) comparing future cases to controls for each study

[c] Information on type of diabetes in family members or other family structure details (i.e. adopted or biological parents) was not collected

[d] Number of biological siblings, including any deceased but not half siblings, was collected in 1996

Table 2 Variance in T2D risk explained by parental history of diabetes (FH) and known genetic (G1) and environmental (E1) risk factors

Model	NHS[a]		HPFS[a]	
	σ^2	95 % CI	σ^2	95 % CI
FH	0.069	0.051–0.087	0.062	0.042–0.082
G1	0.032	0.020–0.044	0.031	0.017–0.045
E1	0.179	0.154–0.204	0.185	0.155–0.215
FH, G1	0.093	0.073–0.113	0.084	0.062–0.106
FH, E1	0.222	0.196–0.248	0.226	0.195–0.257
G1, E1	0.207	0.181–0.233	0.212	0.181–0.243
FH, G1, E1	0.244	0.217–0.270	0.247	0.216–0.278

[a] Observed-scaled σ^2 from linear regressions of FH, G1 (T2D GRS) and E1 (age, BMI, smoking status, diet quality, alcohol consumption and physical activity) and their combinations (specified in column 1) on T2D case-status

obtained estimates of $\hat{\xi} = 0.23$ (95 % CI 0.15–0.27) and $\hat{\pi}_{\mathrm{FH},E} = 0.32$ (95 % CI 0.24–0.36). Setting $x = 0$ (an unrealistically low value) yields only slightly higher estimates of $\hat{\xi}$ (Fig. 1) and $\hat{\pi}_{\mathrm{FH},E}$ for both cohorts.

Our estimate of parental shared environment ($\hat{\xi}$, 23 % where $x = 0.5$) in NHS and HPFS is much lower than that reported in population-based twin studies (Kaprio et al. 1992; Poulsen et al. 1999). In an earlier Finnish Twin study by Kaprio et al. (1992), additive genetics, shared environment and non-shared environment accounted for 42, 30 and 28 % of the variance in T2D, respectively. Corresponding estimates in a recent analysis of the Danish Twin Registry were 26, 41 and 33 % (Poulsen et al. 1999). When adjusting our simulation parameter estimate of G1 from 0.40 to 0.26 [reported by Poulsen et al. (1999)], $\hat{\xi}$ was 29 % in both the NHS and HPFS. In our simulations where estimates of the heritability of T2D range from 20 to 60 % an approximate linear decrease in $\hat{\xi}$ is observed for both cohorts (Fig. 2), and thus remained relatively low in comparison to twin studies. Varying the proportion of E1 explained by genetics (x) had little impact on $\hat{\xi}$, with values ranging from 0.27 to 0.19 across cohorts (Fig. 3).

Supplementary Figure S1 displays results of simulations when extending our definition of FH to include diabetes history of one to five siblings. For both studies, $\hat{\xi}$slightly decreases with each additional sibling. Simulating three first-degree relatives, we repeated the simulation experiment outlined for parental history above for a family history of the disease. When the proportion of E1 explained by genetics (x) is 0.5, the estimate of $\hat{\xi}$ that best fits the observed data in Table 2 (parental or sibling history of diabetes) for NHS was 0.16 (95 % CI 0.11–0.22) and $\hat{\pi}_{\mathrm{FH},E} = 0.25$ (95 % CI 0.19–0.29) (Supplementary Fig. S2a). In HPFS, corresponding estimates of $\hat{\xi}$ and $\hat{\pi}_{\mathrm{FH},E}$ were 0.13 (95 % CI 0.03–0.17) and $\hat{\pi}_{\mathrm{FH},E} = 0.21$ (95 % CI 0.06–0.28), respectively (Supplementary Fig. S2b). Similar patterns to that described for parental history when varying estimates of heritability (G1 + G2) and proportion of E1 explained by genetics (x) were observed when corresponding family history simulations were performed (Supplementary Figs. S2–S4).

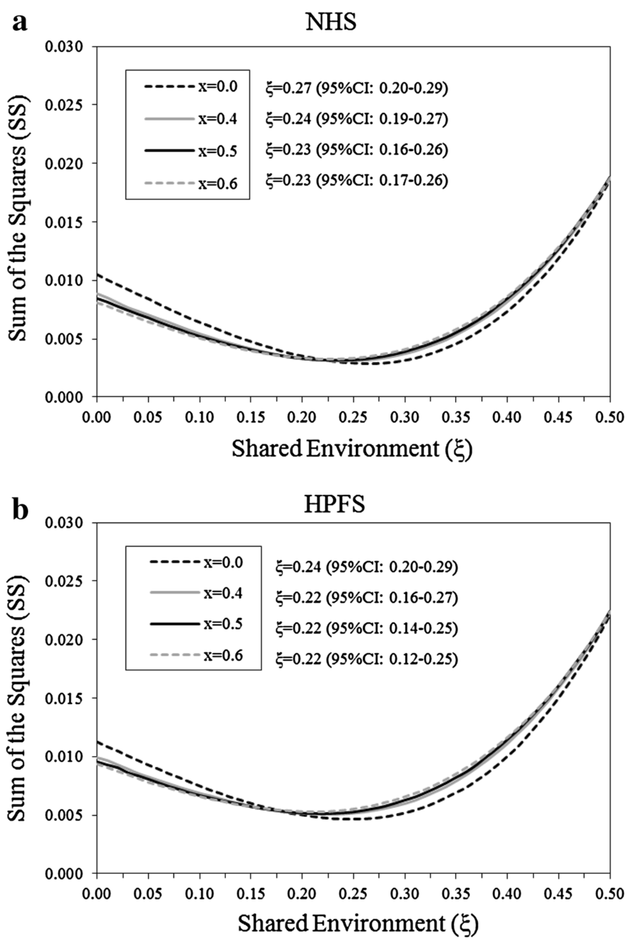

Fig. 1 Estimates of parental shared environment ($\hat{\xi}$). Results for NHS (**a**) and HPFS (**b**) under simulation models, where the heritability of type 2 diabetes ($\sigma^2_{G1} + \sigma^2_{G2}$) is 0.40, $\sigma^2_{G1} = 0.03$, $\sigma^2_{E1} = 0.18$, $\sigma^2_{G2} = 0.37$, $\sigma^2_{E2} = 0.42$ and $x = 0.0$, 0.4, 0.5 or 0.6. The best estimate (or fit) of ξ for a dataset ($\hat{\xi}$) is the value of ξ that minimizes the sum of squared residuals based on differences between observed variances in analyses of real phenotypes and the fitted variances in analyses of simulated phenotypes (see "Materials and methods")

Discussion

In the current study we quantify the extent to which family history derives from genetic or environmental factors. We developed a statistical framework that models four components of variance, including known and unknown genetic and environmental factors, using a liability threshold model. Our approach relies on a comparison between simulations and empirical data. By comparing the association of parental history with T2D in our simulations to case–control studies of T2D nested in the NHS and HPFS, we estimate that first-degree relatives have a correlation of 23 % (95 % CI 15–27 %) in their environmental contribution to T2D liability and that this shared environment is responsible for 32 % (95 % CI 24–36 %) of the association between parental history and T2D, with the remainder due to shared genetics. These findings corroborate and extend the earlier work by Khoury et al. (1988), which relied exclusively on simulations and did not analyze real phenotypes. Although the current study focused on parental history, our framework can also be extended to definitions of family history that include sibling history. Our methods can also be applied to full cohort studies instead of ascertained case–control studies.

Studies of related individuals have traditionally been used to estimate the environmental and genetic components underlying familial aggregation of disease but are often underpowered and subject to a number of biases (Hopper et al. 2005; Kendler et al. 1993; Zaitlen et al. 2013; Zuk et al. 2012). For example, the assumption of equal shared environments between monozygotic and dizygotic twins is often not valid and could lead to an overestimate of shared genetics. The assumption that the correlation of adoptees with their biological parents is due to genetic factors and correlation with their adoptive parents is due to environmental factors may be violated when placement is selective. Estimates of heritability may also be upwardly biased when cases are selected from samples with closely related pairs of individuals (Hopper et al. 2005; Zaitlen et al. 2013) or when certain types of epistatic (gene–gene) interactions exist (Kendler et al. 1993; Zuk et al. 2012). Finally, several conceptual issues surrounding these designs also impact the interpretation and generalizability of the findings.

Studies of unrelated individuals with available genetic, environmental and family history data are less susceptible to the aforementioned biases and their large size, availability and diversity allow greater power. To our knowledge, no previous studies have thoroughly exploited these advantages. Chatterjee et al. (2006) developed a combined approach of kin-cohort and case–control analysis with application to breast cancer to estimate residual familial aggregation, defined as risk conferred by a positive family history excluding risk caused by known variants, but do not partition this estimate into environmental and unknown genetic factors. Bajdik et al. (2001) created a model to simulate the incidence of breast/ovarian cancer in the family of a mutation carrier, but do not incorporate environmental factors. Our approach is based on the liability threshold model (Falconer 1967), which is widely used (Lee et al. 2011; Plomin et al. 2009) but largely untestable with the data we analyze here; we caution that our results could be affected by model mis-specification if this is not the correct model.

The quality and quantity of data collected for a population will be an important determinant of how well our method performs. Although our method is flexible enough to be applied to various case–control settings, the results remain largely population specific. The high level of

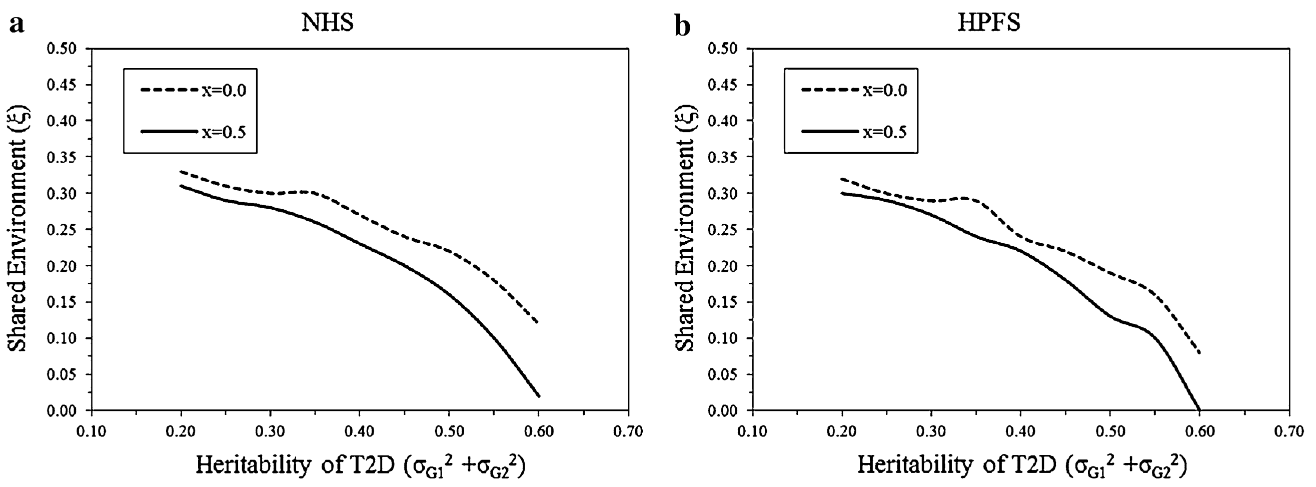

Fig. 2 Parental shared environment ($\hat{\xi}$) under simulation models with varying heritability estimates of type 2 diabetes. Heritability estimates range from 0.20 to 0.60 with increments of 0.05. σ^2_{G1} and σ^2_{E1} are held constant with values of 0.03 and 0.18, respectively. Results for NHS (**a**) and HPFS (**b**)

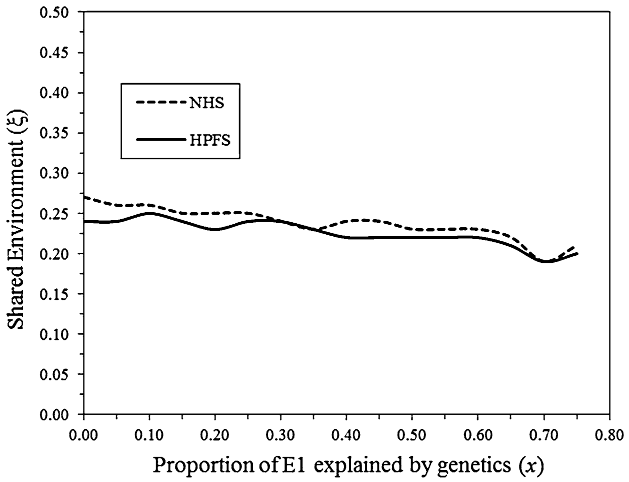

Fig. 3 Parental shared environment ($\hat{\xi}$) under simulation models with varying proportions of E1 explained by genetics (x). $\sigma^2_{G1} = 0.03$, $\sigma^2_{E1} = 0.18$, $\sigma^2_{G2} = 0.37$, $\sigma^2_{E2} = 0.42$. E1 is partitioned into a genetic component E1G ($\sigma^2_{E1G} = \sigma^2_{E1}(x)$) and an environmental component E1E ($\sigma^2_{E1E} = \sigma^2_{E1} - \sigma^2_{E1G}$), where x ranges from 0.0 to 0.75 with 0.05 increments

education and interest in health in the NHS and HPFS may reduce the variance in environmental exposures predisposing to disease which may potentially lead to an overestimation of the genetic contribution to the analysis. The phenotypic variance for height and BMI explained by additive SNP effects in an extended sample of these two cohorts is 35 and 22 %, respectively, and is comparable to those estimated for cohorts with supposedly more generalizable characteristics (Yang et al. 2010a, b), suggesting selection biases in NHS and HPFS may be limited. Like many epidemiological cohorts, the exposure data for NHS and HPFS were largely collected via self-report, which is prone to error and biases and often limits the amount of detailed data collected. In the current study, we used simple definitions of 'parental' and 'family' history as these are frequently used in large population settings. More complex family history 'scores', incorporating age of disease onset, gender and family structures have been proposed (Feng et al. 2009) and might be considered in future applications of our method. Epigenetics is not specifically accounted for in our framework but will be only one of many factors contributing to shared and unshared environmental components. In contrast to 'environment' data, genetic data are measured very well. For T2D, however, the variance explained by known variants (G1) is but a small fraction of estimated heritability (Morris et al. 2012). As more disease loci are discovered (i.e. G1), the performance of our method is likely to improve.

Several risk prediction algorithms that account for family history have been developed (Do et al. 2012; Lewis et al. 2007; Ruderfer et al. 2010; So et al. 2011). Some of these may extend to inclusion of environmental or clinical factors but often assume independence of these risk factors with known genetic variants and family history. Recent work by Chatterjee et al. (2013) has quantified the amount by which family history can improve polygenic risk prediction. However, that study assumed that family history is entirely genetic, whereas our results show that family history is partly environmental, implying that the amount by which family history can improve polygenic risk prediction may be greater than that reported by Chatterjee et al. (2013).

Our findings, if generalizable to other populations, thus inform the interpretation of a 'positive' family history and, consequently, the performance and/or interpretation of studies that seek to use family history as a novel research tool (Ghosh et al. 2012; Thornton and McPeek 2007).

Acknowledgments This work was supported by the National Institute of Environmental Health (R21ES020754-01). The Nurses' Health Study/Health Professionals Follow-up Study Type 2 Diabetes GWAS (U01HG004399) is a component of a collaborative project that includes 13 other GWAS funded as part of the Gene-Environment Association Studies (GENEVA) under the National Institutes of Health (NIH) Genes, Environment, and Health Initiative. Additional funding was provided by the National Cancer Institute (P01CA087969, P01CA055075) and the National Institute of Diabetes and Digestive and Kidney Diseases (R01DK058845).

References

Abbasi A, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AM, van der AD, Navis G, Bakker SJ, Beulens JW (2011) Maternal and paternal transmission of type 2 diabetes: influence of diet, lifestyle and adiposity. J Intern Med 270:388–396

Bajdik CD, Raboud JM, Schechter MT, McGillivray BC, Gallagher RP (2001) A computer model to simulate family history of breast/ovarian cancer in BRCA1 mutation carriers. Math Biosci 171:99–111

Chatterjee N, Kalaylioglu Z, Shih JH, Gail MH (2006) Case–control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. Biometrics 62:36–48

Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet 45:400–405

Chiuve SE, Fung TT, Rimm EB, Hu FB, McCullough ML, Wang M, Stampfer MJ, Willett WC (2012) Alternative dietary indices both strongly predict risk of chronic disease. J Nutr 142:1009–1018

Chu NF, Spiegelman D, Yu J, Rifai N, Hotamisligil GS, Rimm EB (2001) Plasma leptin concentrations and four-year weight gain among US men. Int J Obes Relat Metab Disord 25:346–353

Colditz GA, Hankinson SE (2005) The Nurses' Health Study: lifestyle and health among women. Nat Rev Cancer 5:388–396

Colditz GA, Manson JE, Hankinson SE (1997) The Nurses' Health Study: 20-year contribution to the understanding of health among women. J Womens Health 6:49–62

Cornelis MC, Qi L, Zhang C, Kraft P, Manson J, Cai T, Hunter DJ, Hu FB (2009) Joint effects of common genetic variants on the risk for type 2 diabetes in US men and women of European ancestry. Ann Intern Med 150:541–550

Do CB, Hinds DA, Francke U, Eriksson N (2012) Comparison of family history and SNPs for predicting risk of complex disease. PLoS Genet 8:e1002973. doi:10.1371/journal.pgen.1002973

Falconer DS (1967) The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. Ann Hum Genet 31:1–20

Feng R, McClure LA, Tiwari HK, Howard G (2009) A new estimate of family disease history providing improved prediction of disease risks. Stat Med 28:1269–1283

Garner C (2007) Upward bias in odds ratio estimates from genome-wide association studies. Genet Epidemiol 31:288–295

Ghosh A, Hartge P, Purdue MP, Chanock SJ, Amundadottir L, Wang Z, Wentzensen N, Chatterjee N, Wacholder S (2012) Assessing disease risk in genome-wide association studies using family history. Epidemiology 23:616–622

Hemminki K, Li X, Sundquist K, Sundquist J (2010) Familial risks for type 2 diabetes in Sweden. Diabetes Care 33:293–297

Hopper JL, Bishop DT, Easton DF (2005) Population-based family studies in genetic epidemiology. Lancet 366:1397–1406

Hu FB (2011) Globalization of diabetes: the role of diet, lifestyle, and genes. Diabetes Care 34:1249–1257

Hu FB, Leitzmann MF, Stampfer MJ, Colditz GA, Willett WC, Rimm EB (2001) Physical activity and television watching in relation to risk for type 2 diabetes mellitus in men. Arch Intern Med 161:1542–1548

International Diabetes Federation (2011) IDF diabetes atlas, 5th edn. International Diabetes Federation, Brussels

Kaprio J, Tuomilehto J, Koskenvuo M, Romanov K, Reunanen A, Eriksson J, Stengard J, Kesaniemi YA (1992) Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. Diabetologia 35:1060–1067

Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ (1993) A test of the equal-environment assumption in twin studies of psychiatric illness. Behav Genet 23:21–27

Khoury MJ, Beaty TH, Liang KY (1988) Can familial aggregation of disease be explained by familial aggregation of environmental risk factors? Am J Epidemiol 127:674–683

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34:591–602

Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88:294–305

Lewis CM, Whitwell SC, Forbes A, Sanderson J, Mathew CG, Marteau TM (2007) Estimating risks of common complex diseases across genetic and environmental factors: the example of Crohn disease. J Med Genet 44:689–694

Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, Berglund G, Altshuler D, Nilsson P, Groop L (2008) Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 359:2220–2232

Maes HH, Neale MC, Eaves LJ (1997) Genetic and environmental factors in relative body weight and human adiposity. Behav Genet 27:325–351

Manson JE, Rimm EB, Stampfer MJ, Colditz GA, Willett WC, Krolewski AS, Rosner B, Hennekens CH, Speizer FE (1991) Physical activity and incidence of non-insulin-dependent diabetes mellitus in women. Lancet 338:774–778

Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB Sr, Cupples LA (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359:2208–2219

Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, Strawbridge RJ, Khan H, Grallert H, Mahajan A, Prokopenko I, Kang HM, Dina C, Esko T, Fraser RM, Kanoni S, Kumar A, Lagou V, Langenberg C, Luan J, Lindgren CM, Muller-Nurasyid M, Pechlivanis S, Rayner NW, Scott LJ, Wiltshire S, Yengo L, Kinnunen L, Rossin EJ, Raychaudhuri S, Johnson AD, Dimas AS, Loos RJ, Vedantam S, Chen H, Florez JC, Fox C, Liu CT, Rybin D, Couper DJ, Kao WH, Li M, Cornelis MC, Kraft P, Sun Q, van Dam RM, Stringham HM, Chines PS, Fischer K, Fontanillas P, Holmen OL, Hunt SE, Jackson AU, Kong A, Lawrence R, Meyer J, Perry JR, Platou CG, Potter S, Rehnberg E, Robertson N, Sivapalaratnam S, Stancakova A, Stirrups K, Thorleifsson G, Tikkanen E, Wood AR, Almgren P, Atalay M, Benediktsson R, Bonnycastle LL, Burtt N, Carey J, Charpentier

G, Crenshaw AT, Doney AS, Dorkhan M, Edkins S, Emilsson V, Eury E, Forsen T, Gertow K, Gigante B, Grant GB, Groves CJ, Guiducci C, Herder C, Hreidarsson AB, Hui J, James A, Jonsson A, Rathmann W, Klopp N, Kravic J, Krjutskov K, Langford C, Leander K, Lindholm E, Lobbens S, Mannisto S et al (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44:981–990

Plomin R, Haworth CM, Davis OS (2009) Common disorders are quantitative traits. Nat Rev Genet 10:872–878

Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H (1999) Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance—a population-based twin study. Diabetologia 42:139–145

Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, Pankow JS, Dupuis J, Florez JC, Fox CS, Pare G, Sun Q, Girman CJ, Laurie CC, Mirel DB, Manolio TA, Chasman DI, Boerwinkle E, Ridker PM, Hunter DJ, Meigs JB, Lee CH, Hu FB, van Dam RM (2010) Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet 19:2706–2715

Rice T, Bouchard C, Perusse L, Rao DC (1995) Familial clustering of multiple measures of adiposity and fat distribution in the Quebec Family Study: a trivariate analysis of percent body fat, body mass index, and trunk-to-extremity skinfold ratio. Int J Obes Relat Metab Disord 19:902–908

Rimm EB, Giovannucci EL, Willett WC, Colditz GA, Ascherio A, Rosner B, Stampfer MJ (1991) Prospective study of alcohol consumption and risk of coronary disease in men. Lancet 338:464–468

Ruderfer DM, Korn J, Purcell SM (2010) Family-based genetic risk prediction of multifactorial disease. Genome Med 2:2

So HC, Kwan JS, Cherny SS, Sham PC (2011) Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am J Hum Genet 88:548–565

The InterAct Consortium (2013) The link between family history and risk of type 2 diabetes is not explained by anthropometric, lifestyle or genetic risk factors: the EPIC-InterAct study. Diabetologia 56:60–69

Thornton T, McPeek MS (2007) Case–control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet 81:321–337

van't Riet E, Dekker JM, Sun Q, Nijpels G, Hu FB, van Dam RM (2010) Role of adiposity and lifestyle in the relationship between family history of diabetes and 20-year incidence of type 2 diabetes in US women. Diabetes Care 33:763–767

Vassy JL, Shrader P, Jonsson A, Fox CS, Lyssenko V, Isomaa B, Groop L, Meigs JB, Franks PW (2011) Association between parental history of diabetes and type 2 diabetes genetic risk scores in the PPP-Botnia and Framingham Offspring Studies. Diabetes Res Clin Pract 93:e76–e79

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24. doi:10.1016/j.ajhg.2011.11.029

Willett WC (1998) Nutritional epidemiology. Oxford University Press, New York

Wolf AM, Hunter DJ, Colditz GA, Manson JE, Stampfer MJ, Corsano KA, Rosner B, Kriska A, Willett WC (1994) Reproducibility and validity of a self-administered physical activity questionnaire. Int J Epidemiol 23:991–999

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010a) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569. doi:10.1038/ng.608

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM (2010b) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43:519–525

Zaitlen N, Kraft P, Patterson N, Pasaniuc B, Bhatia G, Pollack S, Price AL (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet 9:e1003520

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci USA 109:1193–1198