Widespread covariation of early environmental exposures and trait-associated polygenic variation

E. Krapohl1, a, L. J. Hanniganb, J.-B. Pingaultc,d, H. Patelc,d, N. Kadeva3, C. Curtisd, G. Breen3,5, S. J. Newhouse3,5, a, T. C. Eley2, F. F. O’Reilly7,3, and R. Plomin1, a

1MRC Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology, and Neuroscience, King’s College London, London SE5 8AF, United Kingdom; 2Clinical, Educational, and Health Psychology, University College London, London WC1E 6BT, United Kingdom; 3Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London SE5 8AF, United Kingdom; 4National Institute for Health Research Biomedical Research Centre, South London and Maudsley National Health Service Foundation Trust, King’s College London, London SE5 8AF, United Kingdom; and 5Farr Institute of Health Informatics Research, UCL Institute of Health Informatics, University College London, London NW1 2DA, United Kingdom

Edited by Christopher F. Chabris, Geisinger Health System, Lewisburg, PA, and accepted by Editorial Board Member Michael S. Gazzaniga August 28, 2017 (received for review May 1, 2017)

Although gene–environment correlation is recognized and investigated by family studies and recently by SNP-heritability studies, the possibility that genetic effects on traits capture environmental risk factors or protective factors has been neglected by polygenic prediction models. We investigated covariation between trait-associated polygenic variation identified by genome-wide association studies (GWAs) and specific environmental exposures, controlling for overall genetic relatedness using a genomic relatedness matrix restricted maximum-likelihood model. In a UK-representative sample (n = 6,710), we find widespread covariation between offspring trait-associated polygenic variation and parental behavior and characteristics relevant to children’s developmental outcomes—Independently of population stratification. For instance, offspring genetic risk for schizophrenia was associated with paternal age (R² = 0.002; P = 1e-04), and offspring education-associated variation was associated with variance in breastfeeding (R² = 0.021; P = 7e-30), maternal smoking during pregnancy (R² = 0.008; P = 5e-13), parental smoking (R² = 0.01; P = 4e-15), household income (R² = 0.032; P = 1e-22), watching television (R² = 0.034; P = 5e-47), and maternal education (R² = 0.065; P = 3e-96). Education-associated polygenic variation also captured covariation between environmental exposures and children’s attention/hyperactivity, conduct problems, and educational achievement. The finding that genetic variation identified by trait GWASs partially captures environmental risk factors or protective factors has direct implications for risk prediction models and the interpretation of GWAS findings.

environmental risk | polygenic scores | gene–environment correlation | human complex traits | prediction

Environmental exposures are among the best early predictors of developmental outcomes. For instance, maternal smoking during pregnancy, socioeconomic deprivation, and time spent watching television and playing video games are associated with lower academic achievement (1–9). Harsh paternal physical discipline such as hitting has been linked to increased emotional and behavioral problems including aggression in adolescence (10–14). Paternal age is a risk factor for a range of disorders and subclinical phenotypes including low academic achievement (15), with the link to autism spectrum disorders and schizophrenia most robustly replicated (16–21). Breastfeeding and higher parental socioeconomic status (education, income, occupation) are protective factors for a range of outcomes including educational achievement (7, 8, 22).

Evidence from many family, twin, and adoption studies converges in showing that individuals’ exposure to environments partially depends on their genotype (i.e., genotype–environment correlation). This includes both parenting characteristics and broad socioeconomic variables; all are partially heritable (23–28). In the past decade, quantitative genetic research of this type has been extended to explore genetic and environmental contributions to correlations between environmental factors and children’s outcomes (29–32). Some new designs such as the children-of-twins designs make it possible to tease apart different types of genotype–environment correlation and identify environmental influences free of genetic confounds (33–37). These designs are limited by the extent to which environmental variables differ between close relatives.

Converging evidence for gene–environment correlation comes more recently from “single nucleotide polymorphism (SNP)-heritability” studies that estimate overall genetic influences from genome-wide DNA differences in unrelated individuals. These studies have shown that variation in individuals’ social deprivation, household income, stressful life events, and family socioeconomic status partially reflects individuals’ differences across genome-wide common genetic variants measured on SNP arrays (38–44). There have also been a few reports of extending SNP heritability analysis to estimate genetic correlations between environmental measures and measures of children’s developmental outcomes (38–40).

Gene–environment correlation is recognized and investigated by family studies and recently by SNP-heritability studies. However, the possibility that genetic effects on traits capture environmental risk factors or protective factors has been neglected by polygenic prediction models, which use trait-associated genetic

Significance

Environmental exposures are among the best predictors of health and educational outcomes. Models that estimate the effect of environmental exposures on developmental outcomes typically ignore genetic factors or focus on gene–environment interaction (whether individuals’ response to environmental exposures depends on their genotype). Here we test gene–environment correlation (whether individuals’ exposure to environments depends on their genotype). Using a method that tests specific genetic effects while controlling for background genetic effects, we estimate covariation between children’s genetic liability/propensity for core developmental outcomes and a wide range of environmental exposures. Findings suggest that genetic variants associated with traits, such as educational attainment, body mass index, and schizophrenia, also capture environmental risk and protective factors.

Author contributions: E.K. and R.P. designed research; E.K. performed research; E.K., L.J.H., H.P., N.K., C.C., G.B., and S.J.N. contributed new reagents/analytic tools; E.K. analyzed data; and E.K., L.J.H., J.-B.P., T.C.E., P.F.O., and R.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. C.F.C. is a guest editor invited by the Editorial Board.

1To whom correspondence may be addressed. Email: eva.krapohl@kcl.ac.uk or robert.plomin@kcl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1707178114/-/DCSupplemental.
variants identified by genome-wide association studies (GWASs) to estimate individual-level genetic trait propensities for trait prediction.

Here we tested whether genetic variation identified by trait GWASs captures variation in environmental risk factors or protective factors. Specifically, as children’s environments and genetic propensities are both “provided by” their parents, these are expected to correlate because parents pass on genetic variants to their offspring that influence parents’ environment-providing behaviors. Therefore, we examine to what extent offspring trait-associated alleles covary with parental traits and behaviors previously reported to be environmental risk or protective factors for important child outcomes. We also tested to what extent offspring genetic trait propensities contribute to the correlation between parenting characteristics and children’s developmental outcomes.

First, we conducted a systematic investigation of covariation between children’s genetic propensities for specific developmental outcomes and a wide range of environmental exposures previously shown to be risk or protective factors for these outcomes (SI Appendix, Methods S3). We focus on genetic propensities—that is, individual-specific genomic profiles of trait-associated alleles—for three core developmental outcomes: educational attainment (45), body mass index (BMI) (46), and schizophrenia (47). These traits form three important domains of child development—social-cognitive, mental health, and physical health—each are robust predictors of mortality and life expectancy, with substantial associated societal and personal burden (48–55). They were chosen because of the availability of statistically powerful GWAS summary statistics for these traits (56).

Second, we tested whether the environmental exposures predicted children’s developmental outcomes (as would be expected based on previous literature) and to what extent these associations are captured by children’s polygenic propensities for education, BMI, and schizophrenia. For this, we examined associations between the environmental exposures and three developmental outcomes assessed at age 16 in our sample: educational achievement, inattention-hyperactivity symptoms, and conduct problems (SI Appendix, Methods S3).

We used a sample of 6,710 unrelated individuals, drawn from the Twins Early Development Study (TEDS), for whom genotype data and a wide range of specific environmental exposure measures and developmental outcomes from birth to adolescence are available. TEDS is a multivariate longitudinal study that recruited over 11,000 twin pairs born in England and Wales in 1994, 1995, and 1996 (57, 58), shown to be representative of the UK population (38, 59).

We created genome-wide polygenic scores for trait-associated genetic variants for each individual in the sample using summary statistics from the independent GWASs of years of education (EDU) (45), BMI (46), and schizophrenia (SCZ) (47). We used a Bayesian approach (60) that estimates posterior mean effect sizes of each marker by using a point-normal mixture prior on a Bayesian approach (60) that estimates posterior mean effect statistics from the independent GWAS of years of education (EDU) (45), BMI (46), and schizophrenia (SCZ) (47). We used a Bayesian approach (60) that estimates posterior mean effect sizes of each marker by using a point-normal mixture prior on a Bayesian approach (60) that estimates posterior mean effect statistics from the independent GWAS of years of education (EDU) (45), BMI (46), and schizophrenia (SCZ) (47). We used a Bayesian approach (60) that estimates posterior mean effect sizes of each marker by using a point-normal mixture prior on a Bayesian approach (60) that estimates posterior mean effect statistics from the independent GWAS of years of education (EDU) (45), BMI (46), and schizophrenia (SCZ) (47).

Results

To estimate the univariate effect of each polygenic score on the environmental exposures, we fit a series of single-score models, which reveal significant trait-associated polygenic effects across a wide range of environmental exposures. Fig. 1, Left (and SI Appendix, Table S1) shows the estimated variance explained by each polygenic score for each of the environmental measures. Environmental factors varied significantly as a function of trait-associated polygenic variation, independently of population stratification. This provides evidence for trait-associated genotype–environment correlation. However, given the robust evidence for extensive pleiotropy across complex traits (61), we aimed to isolate the effects of each trait-associated polygenic score using a multiscore model. To test the trait specificity of the polygenic effects on environmental exposures, we jointly modeled the three scores for years of education, BMI, and schizophrenia, allowing us to estimate the effects of each polygenic score while statistically adjusting for the effects of the others. Fig. 1, Right (and SI Appendix, Table S2) shows that the multiscore models revealed some attenuation of the polygenic score effects compared with the single-score models, suggesting that the effects of the three scores on environmental exposures are nonindependent. Specifically, the effects of BMI-associated polygenic variation on several environmental measures (including watching television and parental education) were no longer significant.

Breastfeeding duration was positively associated with offspring education polygenic score, adjusted for BMI and schizophrenia polygenic scores ($R^2 = 0.021$, beta = 0.144; $P = 7e-30$). Fig. 2A displays children’s adjusted education polygenic score as a function of whether and for how long they were breastfed. Children who were breastfed had, on average, an education polygenic score approximately one quarter SD lower (Hedges’ $g = 0.26$) than children whose mothers did not smoke ($t = 7.93$, df = 1,556; $P = 4e-15$), and 3.4% in television watching in the household (beta = 0.184, $P = 5e-47$).

Offspring genetic risk for schizophrenia was positively associated with paternal age, even when adjusting for education and BMI-associated polygenic variation ($R^2 = 0.002$, beta = 0.049; $P = 1e-04$). Fig. 2C shows children’s adjusted genetic risk for schizophrenia as a function of paternal age. Children whose father was aged over 45 at their birth had, on average, a genetic risk score for schizophrenia over one quarter SD (Hedges’ $g = 0.26$) higher than children whose father was under the age of 26 at their birth ($t = 3.01$, df = 411; $P = 3e-03$).

Next, we examined the extent to which associations between environmental exposures and developmental outcomes were explained by trait-associated polygenic variation for education, BMI, and schizophrenia (SI Appendix, Fig. S3). We examined associations between environmental exposures and three developmental outcomes: educational achievement, inattention-hyperactivity symptoms, and conduct problems. Of the three polygenic scores, only the education polygenic score captured covariation between environmental exposures and the three developmental outcomes (SI Appendix, Table S3).

On average education-associated polygenic variation explained 15% of the associations between the environmental measures and children’s developmental outcomes. For example, the education polygenic score explained 23% ($P = 1.2e-18$) of the beta = 0.19 covariance between child educational achievement and breastfeeding. Education-associated polygenic variation also captured 17% ($P = 4.4e-08$) of the association between parental slapping/smacking and conduct problems and hyperactivity/inattention problems (beta = 0.20 for both).

Discussion

We report evidence for covariation between trait-associated polygenic variation and early environmental exposures independently
of population stratification. We show that a wide range of parental, neighborhood, and parent–child perinatal characteristics, representing key early life “environmental” influences, present at birth or early in life, correlate with offspring genetic propensity—specifically, with the allele frequency at loci associated with education, BMI, and schizophrenia. We also demonstrate that covariance between environments and important developmental outcomes are partially captured by education-associated polygenic variation.

The present study combines family and molecular data. In addition to replicating the general finding that individuals’ environmental exposures vary as a function of their genotype, the current findings suggest that trait GWASs are detecting genetic variants associated with parental characteristics and their correlation with child outcomes.

Importantly, the association between exposures and outcomes was by no means entirely captured by offspring trait-associated polygenic variation. There are three likely, nonmutually exclusive, explanations for this. First, a substantial proportion of the exposure–outcome associations is likely due to nongenetic factors. Second, polygenic scores intrinsically underestimate the total genetic effects on the exposure–outcome associations because they are limited to the additive effects of common variants on a particular trait that the discovery GWAS was powered to detect. Third, we only measure offspring polygenic variation, but offspring phenotype can be influenced not only by transmitted but also by nontransmitted parental alleles via parental phenotype (i.e., child exposure).

The education-associated polygenic variation showed the strongest and most consistent correlations with environmental exposures. This is consistent with research showing associations between educational attainment and many parental behaviors and characteristics (e.g., refs. 12, 31, 62, and 63). Moreover, the multipolygenic score models showed that the association between BMI-associated polygenic variation and environmental exposures such as television watching and parental education is explained by education-associated genetic variations. This suggests the potential for multipolygenic models for isolating polygenic effects, provided the underlying discovery GWASs are similarly powered. The finding of an association between paternal age and offspring genetic risk for schizophrenia is consistent with previous evidence for older fathers’ elevated risk for conceiving a child who will go on to develop schizophrenia (18, 19, 63). Although the current findings provide evidence for the relevance of gene–environment correlation for polygenic trait prediction methods, they are not informative about the mechanisms involved.

The observed associations could arise from passive or active gene–environment correlation or via environmentally mediated genetic effects, all of which are nonmutually exclusive. Fig. 3 illustrates these possibilities schematically. Many of the observed associations between offspring genotype and environment-providing parental characteristics are outside of the offspring’s influence (e.g., parental age and education level at child birth) and are therefore likely to result from passive gene–environment correlation. That is, parental genetic propensities that were passed down to offspring are also associated with environment-providing parental behavior (through both path a and b, Fig. 3). However, some of the investigated parental behaviors could partially be evoked by offspring genetic propensities (through paths c and d in Fig. 3; e.g., breastfeeding, watching television). Finally, genetic correlations could arise as a result of environmentally mediated genetic effects (e.g., if education-associated genetic variation influenced mothers’ predisposition to smoke during pregnancy and prenatal exposure to nicotine had an environmental effect on offspring attention problems, this could result in offspring education-associated polygenic variation being associated with maternal smoking pregnancy as well as capturing part of its correlation with offspring attention problems).

The design of the current study is unable to distinguish environmentally mediated genetic effects, passive, and evocative gene–environment correlations. One way to investigate the contributions of these different mechanisms would be to use samples incorporating parental genotype data. In analyses of such samples,
Materials and Methods

We used genome-wide SNP and environment-wide phenotype data from 6,710 unrelated individuals drawn from the UK-representative TEDS (57, 58). TEDS data can be accessed in accordance with the Data Access Policy, which can be viewed at www.teds.ac.uk/research/collaborators-and-data/teds-data-access-policy. We processed the 6,710 genotypes using stringent quality control procedures followed by imputation of SNPs to the Haploype Reference Consortium reference panel (65) (SI Appendix, Methods S1). This included removing one individual from any pair of individuals with an estimate SNP marker relatedness > 0.05. After quality control, 7,581,516 genotyped or well-imputed (info > 0.70) variants remained.

Polygenic Scores. For each individual in the sample, we created polygenic scores for years of education, schizophrenia, and BMI. After coordinating overlapping markers between each of the three GWAS summary statistics and the target data by excluding markers due to nucleotide inconsistencies or low minor allele frequency (<1%), we retained 5,690,632 for the years of education (45), 5,781,731 for schizophrenia (47), and 1,810,667 for BMI (46). We constructed polygenic scores as the effect-size weighted sums of individuals’ trait-associated alleles across all SNPs. We used LDpred (60), which places a prior on the markers’ effect sizes and adjusts summary statistics for LD between markers. For each trait, we created the score using three different priors on the fraction of causal markers—0.01, 0.1, and 1.0—from which the one yielding the largest R² in the single-polygenic score models was then entered into the multipolygenic score model. For details on the polygenic score construction, see SI Appendix, Methods S2.

To account for population stratification, we adjusted the polygenic predictors by the first 30 principal components (PCs) generated from genotype data before the analysis. We used the top 30 PCs as well as genotype array and plate to create a N × P matrix Z of eigenvectors across the P selected PCs. We then regressed the genetic polygenic predictor onto the eigenvectors as S = μ + Zβ + e, where μ is the mean and β is a P × 1 vector of the regression coefficients and e is the residual error.

Single-Score and Multiscore Genomic-Relatedness Matrix Restricted Maximum-Likelihood Models. When estimating genetic effects on environmental exposures, the possibility of population stratification is especially salient. This is because genetic and common environment effects, even if uncorrelated, may confound the genetic effects estimates. To account for this, we used the Restricted Maximum Likelihood (REML) method to estimate the genetic and environmental effects while controlling for the genetic and environmental covariates. We estimated the genetic and environmental effects using the following model:

...
be confounded as close relatives share both genes and their environment to a greater extent than other individuals. We control this type of confounding because, under only population stratification, we would not expect an association between polygenic predictors and environmental measures within the mixed effect model of Eqs. 1 and 2. This is because they account for population stratification by both regressing PCs from the polygenic predictors (see above) and fitting a relationship matrix estimated from all SNP markers (see below).

To estimate the degree to which trait-associated polygenic variation captures variation in environmental measures, we estimated the relationship between the polygenic scores and the environmental measures, while controlling for net genetic relatedness by fitting the effects of all of the SNPs as random effects by a mixed linear model:

Single-score model: \( \text{var}(y) = \mu + s_i + \alpha \sigma_i^2 + \epsilon^2 \) \[1\]

Multiscore model: \( \text{var}(y) = \mu + S_{\text{BMI}} + S_{\text{SCZ}} + S_{\text{EDU}} + \alpha \sigma_i^2 + \epsilon^2 \) \[2\]

\( y \) is an \( n \times 1 \) vector containing the level of environmental exposure, with \( n \) being the sample size. \( \beta \) is a vector of fixed effects estimating the effects of the polygenic predictor, independently of overall genetic relatedness \( g \).

In the single-score model (Eq. 1), \( S_i \) is a vector containing individuals’ polygenic score for one of \( i \in [\text{years of education (EDU)}, 45], \text{BMI, 46}, \text{schizophrenia (SCZ), 47] \), adjusted for 30 PCs, genotyping array, and plate (see section above). \( g \) is an \( n \times 1 \) vector of the total genetic effects of the individuals, independently of \( \beta \), with \( g \sim N(0, \Sigma_g^2) \), and \( A \) is interpreted as the genetic relationship matrix (GRM) between individuals. \( \alpha > 0.01 \); relatedness < 0.05 as described above). The genomic relationship of each pair of subjects \( j \) and \( k \) is calculated as \( A_{jk} = 1 - \frac{1}{\text{MD}} \), where \( \text{MD} \) is the number of copies of the reference allele for the \( i \)th SNP in the \( j \)th individual and \( p_i \) being the frequency of the reference allele (66).

In the multiscore model (Eq. 2), the effects of the three polygenic predictors are being estimated jointly, thereby allowing the effect of each polygenic predictor independently of each other and of overall genetic relatedness \( g \).

The genetic relatedness matrix accounts for population stratification in the environmental exposure, because it is equivalent to fitting all of the PCs within the model. Eqs. 1 and 2 were estimated using the restricted maximum likelihood (REML) approach implemented in the reml function in GCTA v1.26.0 (67).

**Decomposition of Covariance Between Environmental Exposures and Developmental Outcomes.** We fit structural equation models to decompose the covariance between environmental exposures and developmental outcomes into effects of the three polygenic scores and residual covariance (SI Appendix, Fig. 3). The total covariance estimated as Cov\(_{\text{total}} = (a \times d) + (b \times e) + (c \times f) + g \) was decomposed into the effects of the education score \( \text{Cov}_{\text{EDU}} = a \times d \), the BMI score \( \text{Cov}_{\text{BMI}} = b \times e \), the schizophrenia score \( \text{Cov}_{\text{SCZ}} = c \times f \), and residual covariance \( g \). We used maximum likelihood estimation with robust (Huber-White) SEs. The analyses were conducted using the lavaan package in R (68).

**Multiple Testing Correction.** \( P \) values obtained for each statistic were corrected for multiple testing using the Sidák correction (69). The Sidák adjusted alpha level is equal to \( 1 - (1 - p^k)^{1/k} \), where \( k \) is the number of tests. The total number of tests was 357, with 153 (3 scores \( \times 3 \) priors \( \times 17 \) exposures) tests for the single-polygenic score models, 51 (3 scores \( \times 17 \) exposures) tests for the multilpolygenic score model, and 153 (3 scores \( \times 17 \) exposures \( \times 3 \) outcomes) test for the decomposition of covariance models. The multiple comparison adjustments were applied to \( \alpha = 0.05 \). Hence, the corrected “experimentwise” \( \alpha \) level was \( 1 - (1 - 0.05)^{357} = 1.43 \times 10^{-4} \).